Graduate Texts in Mathematics




Richard Beals
Roderick S. C. Wong

# Explorations in Complex Functions



Graduate Texts in Mathematics 287

# Graduate Texts in Mathematics



**Graduate Texts in Mathematics** bridge the gap between passive study and creative understanding, offering graduate-level introductions to advanced topics in mathematics. The volumes are carefully written as teaching aids and highlight characteristic features of the theory. Although these books are frequently used as textbooks in graduate courses, they are also suitable for individual study.


More information about this series at http://www.springer.com/series/136

Richard Beals · Roderick S. C. Wong

# Explorations in Complex Functions



Richard Beals
Department of Mathematics
Yale University
New Haven, CT, USA

Roderick S. C. Wong
Department of Mathematics
City University of Hong Kong
Kowloon, Hong Kong

ISSN 0072-5285 ISSN 2197-5612 (electronic)
Graduate Texts in Mathematics
ISBN 978-3-030-54532-1 ISBN 978-3-030-54533-8 (eBook)
https://doi.org/10.1007/978-3-030-54533-8

Mathematics Subject Classification: 30-01, 33-01, 30D35, 33E05, 11M06



This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

A friend of one of the authors has written a book with the (ironic) title *In the Midst of Plenty*. This would be an apt, if unusual, choice of title for this book. A first course in complex analysis introduces keys that unlock many doors. One such key is the residue theorem, in its many forms; another is analytic continuation. The doors open onto many subjects of interest. Too many subjects, in fact, to cover in a single follow-up course.

This book assumes as background a standard first course in complex analysis. Our purpose is to provide relatively brief, but self-contained, introductions to many of the subjects alluded to above. Some of these subjects are within the mainstream of complex analysis itself. Other topics provide tools that are widely used within pure mathematics, or that have applications beyond mathematics, or both. Some topics come up in different contexts in different chapters. For example, there are two proofs of Picard's "little" theorem, and several discussions of the parametrization of algebraic curves.

Chapter 1 is a summary, with selected proofs, of material from a basic course in complex analysis. Included are Cauchy's theorem and consequences (integral formula, series expansion, residue theorem, maximum modulus principle, reflection principle). Also included are the basics of infinite products and analytic continuation. For later use the chapter contains introductions to the Stieltjes integral relative to a jump function, to Hilbert spaces, and to $L^p$ spaces (as completions of spaces of nice functions).

Chapter 2 introduces the Riemann sphere and its automorphism group, the group of linear fractional transformations. The cross product and general mapping properties are covered, and the automorphism groups of the half-plane and the disk are identified. These considerations lead naturally to hyperbolic geometry in the disk or half-plane, which is the subject of Chapter 3.

Chapter 4 introduces harmonic functions in the plane. The Dirichlet problem and Poisson's formula lead to the Weierstrass approximation theorems, and to the Riesz-Fischer theorem for Fourier series. The Schwarz reflection principle prepares the way to the results on boundary behavior of conformal maps that are covered in several later chapters.

Riemann's mapping theorem, the Schwarz–Christoffel formulas, and univalent functions are covered in Chapter 5. In Chapter 6 the Schwarzian derivative is introduced as a measure of curvature, followed by the proof that mappings to curvilinear polygons are quotients of solutions of Fuchsian equations. Particular cases of this are mappings to triangles, or to regular polygons, shown to be quotients of hypergeometric functions.

Chapter 7 covers analytic continuation, Riemann surfaces of functions, algebraic curves, and compact Riemann surfaces. The chapter concludes with a very brief introduction to surfaces of higher genus, exemplified by the Bolza surface. This chapter leads, in a way, to the following two chapters, as well as to later chapters on elliptic functions.

The Weierstrass product theorem and Hadamard's product formula for functions of finite order are the focus of Chapter 8. Application is made to Riemann's xi function, and to an eigenvalue problem. Chapter 9 introduces Nevanlinna's value distribution theory for entire meromorphic functions, starting with Jensen's theorem and the Nevanlinna and Ahlfors–Shimizu characteristics. Nevanlinna's second fundamental theorem is shown to have applications to two theorems of Picard.

Chapter 10 introduces Euler's two definitions of the gamma function, the beta function, the reflection formula, and Legendre's duplication formula. Included is a far-reaching extension, due to Stieltjes, of Stirling's asymptotic approximation, which is important for the study of the Riemann hypothesis in Chapter 13.

Chapters 11 through 13 make up an introduction to analytic number theory. Riemann's zeta function and xi function in Chapter 11 lead to Dirichlet *L*-functions and Dirichlet's theorem on primes in arithmetic progressions in Chapter 12. Chapter 13 treats the prime number theorem in the context of the Riemann hypothesis, and also the relation of the Riemann hypothesis to the accuracy of Gauss's approximate formula for the distribution of primes.

Chapters 14 through 16 introduce elliptic functions and three approaches to their construction. The general theory and construction by means of theta functions, are covered in Chapter 14. Chapters 15 and 16 are independent of each other. The pendulum equation leads to the Jacobi elliptic functions in Chapter 15. Weierstrass's direct construction, starting with the period lattice, is covered in Chapter 16.

The Weierstrass theory of Chapter 16 leads to the study of the modular function, Picard's theorems, and a glance into automorphic functions and the $J$ function in Chapter 17. (An appendix notes the connection to "moonshine" and the monster group.)

Chapter 18 introduces approximate identities, Schwartz functions, and the Cauchy and Hilbert transforms. The Cauchy transform leads naturally to the Fourier transform in $L^1(\mathbb{R})$ and in $L^2(\mathbb{R})$. This, in turn, prepares the way for the following two chapters, which are independent of each other.

Chapter 19 treats the Phragmén–Lindelöf principle. This principle is applied in Hardy's characterization of the Gaussian probability distribution, in the proof of the Paley–Wiener theorem, and in the proof of a theorem of Hardy concerning functions of exponential type.

A theorem of Wiener, and its generalization by Lévy, are proved in Chapter 20. These theorems are used in a version, due to Gohberg and Krein, of the Wiener–Hopf approach to equations of convolution type on the half-line.

Chapters 21 through 23 are generally independent of each other and of earlier chapters (other than Chapter 1). Chapter 21 treats some tauberian theorems, from Tauber and Hardy through Karamata and Wiener. A theorem of Malliavin provides an error estimate that is applicable to distribution of eigenvalues. The section on Wiener's theorem has some dependence on results from Chapter 18.

Chapter 22 introduces the method of steepest descent. Applications include asymptotics of the Airy integral, and the Hardy–Ramanujan theorem on asymptotics of the partition function. Chapter 23 sketches the complex interpolation method, interpolation of $L^p$ spaces, and the Riesz–Thorin theorem, with application to Fourier series.

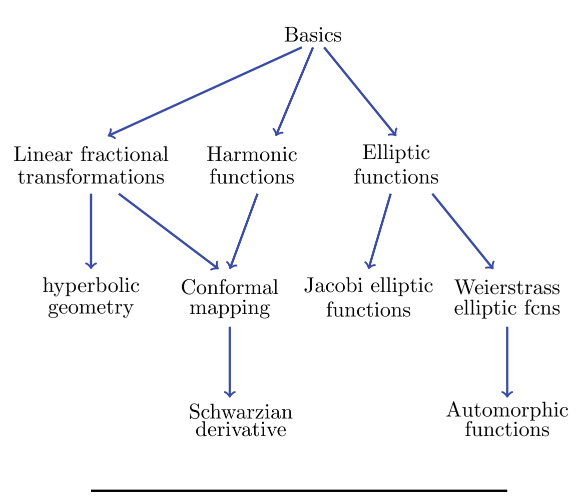
We have tried to make the various presentations self-contained. This has led us to some short excursions into real analysis, functional analysis, and algebra. In particular we have included expositions of irreducibility for polynomials in two variables, and of the character theory of finite abelian groups.

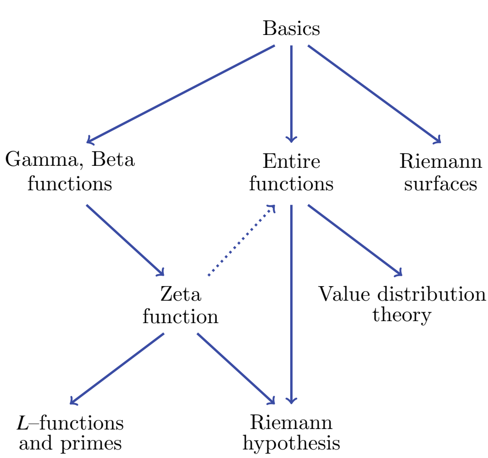
We have also tried to make the various chapters as independent as possible. Charts that show the principal dependence relations follow this preface.

The first author used selections of drafts of a number of these chapters in two versions of a second-semester complex variables class at Yale in 2018 and 2019. He is grateful to the students for their indulgence and attentiveness. Their pertinent questions and comments led to a number of corrections and clarifications. The first author is also grateful to the staff of the Liu Bei Ju Center of City University of Hong Kong, as well as Dr. Huang Xiaomin and Dr. Wang Xiangsheng for their assistance and technical help in June 2019. The second author also thanks Dr. Huang Xiaomin for her considerable help as a postdoc in 2019. Both authors are grateful to Alberto Guzman for critical reading of much of the manuscript, and to their wives, Nancy and Edwina, for their unfailing moral support.

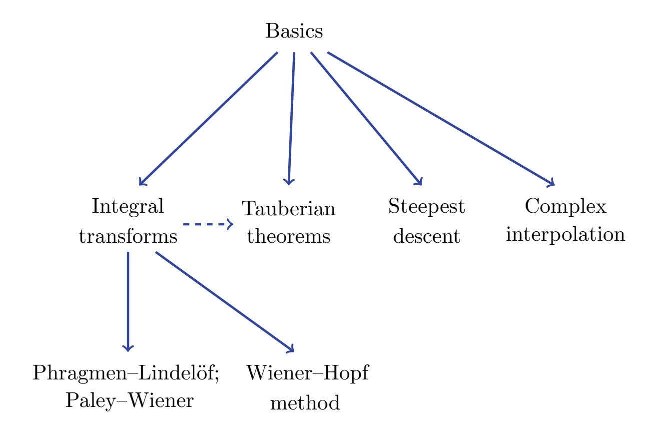
New Haven, CT, USA — Richard Beals
Kowloon, Hong Kong — Roderick S. C. Wong
May 2020

## Dependence relations among chapters: 2 charts



## Dependence relations among chapters: one more chart



# Contents

# Chapter 1
# Basics




This chapter begins with a brief summary of facts from a standard introductory complex variables course: Cauchy's formula and consequences, isolated singularities, residues, and the complex logarithm. Also included are three topics that are not as standard for an elementary course, but are used in many of the following chapters: reflection properties, infinite products, and analytic continuation. For all this material we give brief discussions and sketches of proofs.

The chapter concludes with an outline of basic facts about three subjects from real analysis and functional analysis that occur in one or more later chapters: the Stieltjes integral (discrete case), Hilbert space, and $L^p$ spaces (including duality).

Throughout, a *domain* $\Omega$ is a non-empty open subset of the complex plane $\mathbb{C}$. In this chapter it is assumed that $\Omega$ is bounded and that the boundary $\partial\Omega$ is the union of finitely many piecewise smooth closed curves, oriented so that $\Omega$ lies to the left of each boundary curve. All curves are assumed to be piecewise smooth, and *simple* (having no self-intersections).

## 1.1 The Cauchy–Riemann equations and Cauchy's integral theorem

Real numbers are denoted here by $x$, $y$. Recall that if $z = x + iy$, then $\bar{z} = x - iy$ and

$$\begin{aligned} z = r(\cos\theta + i\sin\theta) \;&=\; re^{i\theta}, \\ r^2 = x^2 + y^2 \;=\; z\bar{z}, \quad \theta \;&=\; \tan^{-1}(y/x) \;=\; \arg z. \end{aligned} \tag{1.1.1}$$

Consider a function

$$f(x+iy) \;=\; u(x,y) + i\,v(x,y), \quad x+iy \in \Omega,$$

where $u$ and $v$ are real-valued and have continuous first partial derivatives. The complex-valued function $f$ is *holomorphic* (differentiable in the complex sense) if



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_1

and only if $u$ and $v$ satisfy the *Cauchy–Riemann equations*:

$$u_x = v_y, \qquad u_y = -v_x, \tag{1.1.2}$$

where the subscripts denote partial differentiation.

Green's theorem (or an argument due to Goursat that uses only pointwise differentiability) yields the basis theorem of the subject.

**Theorem 1.1.1. (Cauchy integral theorem)** *If $f$ is holomorphic in a domain $\Omega$, and continuous on the closure of $\Omega$, then*

$$\int_{\partial\Omega} f(\zeta)\,d\zeta = 0. \tag{1.1.3}$$

Let us pause to look at the Cauchy–Riemann equations and Cauchy's theorem from the point of view of differential forms and Green's theorem. The pairs of 1-forms $dz$, $d\bar{z}$, and $dx$, $dy$ are related by

$$dz = dx + i\,dy, \qquad d\bar{z} = dx - i\,dy;$$
$$dx = \frac{dz + d\bar{z}}{2}, \qquad dy = \frac{dz - d\bar{z}}{2i}.$$

Thus

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy = \frac{1}{2}\left[\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right]dz + \frac{1}{2}\left[\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right]d\bar{z}.$$

It is natural to express this as

$$df = \frac{\partial f}{\partial z}dz + \frac{\partial f}{\partial \bar{z}}d\bar{z} = \partial f\,dz + \bar{\partial} f\,d\bar{z},$$

where

$$\partial = \frac{\partial}{\partial z} = \frac{1}{2}\left[\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right]; \qquad \bar{\partial} = \frac{\partial}{\partial \bar{z}} = \frac{1}{2}\left[\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right]. \tag{1.1.4}$$

With $f = u + iv$ we find that

$$\partial f = \frac{1}{2}\left[(u_x + v_y) + i(v_x - u_y)\right], \qquad \bar{\partial} f = \frac{1}{2}\left[(u_x - v_y) - i(v_x + u_y)\right]. \tag{1.1.5}$$

Thus the Cauchy–Riemann equations (1.1.2) are equivalent to the single equation $\bar{\partial} f = 0$.

A standard form of Green's theorem is that if $\Omega$ is a domain, then

$$\int_{\partial\Omega} [P\,dx + Q\,dy] = \iint_{\Omega} [Q_x - P_y]\,dx\,dy. \tag{1.1.6}$$

It is an exercise, using the identities above, to show that (1.1.6) is equivalent to the equation

$$\int_{\partial\Omega}[f\,dz + g\,d\bar{z}] \;=\; 2i\iint_{\Omega}[\bar{\partial}f - \partial g]\,dx\,dy. \tag{1.1.7}$$

In particular, taking $g = 0$ and assuming that $\bar{\partial}f = 0$, we obtain (1.1.3) as a particular case.

Another application of these identities is the calculation of the *area* of the image of a domain $\Omega$ under an injective holomorphic function $f$ whose first and second partial derivatives are continuous up to the boundary. If $f = u + iv$, the area is

$$\iint_{\Omega}\begin{vmatrix} u_x & v_x \\ u_y & v_y \end{vmatrix}\,dx\,dy \;=\; \iint_{\Omega}[u_x v_y - u_y v_x]\,dx\,dy \;=\; \iint_{\Omega}[u_x^2 + v_x^2]\,dx\,dy.$$

Note that $\partial f = u_x + iv_x$, and $\partial[\bar{f}] = 0$, so the integrand is

$$\partial f\,\overline{\partial f} \;=\; \bar{\partial}[\partial f \cdot \bar{f}] \;=\; \bar{\partial}[f'\bar{f}].$$

It follows from (1.1.7) that for holomorphic injective $f$,

$$\text{Area}\,\{f(\Omega)\} \;=\; \frac{1}{2i}\int_{\partial\Omega} f'(z)\overline{f(z)}\,dz. \tag{1.1.8}$$

## 1.2 The Cauchy integral formula and applications

Much of basic complex function theory consists of exploring (fairly immediate) consequences of Theorem 1.1.1. One such consequence is the *Cauchy integral formula*. If $f$ is holomorphic in a general domain $\Omega$, and continuous on the closure, we can apply (1.1.3) to the function

$$g(w) \;=\; \frac{1}{2\pi i}\cdot\frac{f(w)}{w-z}, \qquad w \in \Omega,$$

on the domain $\Omega_\varepsilon$ formed by removing from $\Omega$ a small disk centered at $z$,

$$D_\varepsilon(z) \;=\; \{w : |w-z| < \varepsilon\} \;=\; \{w : w = z + re^{i\theta},\; r < \varepsilon,\; 0 \le \theta < 2\pi\}$$

The integral over the boundary of $D_\varepsilon$, oriented in the positive (counterclockwise) direction, approaches $f(z)$ as $\varepsilon \to 0$; see the calculation (1.2.4). Taking the limit yields the formula (1.2.1). This formula can be differentiated arbitrarily often.

**Theorem 1.2.1. (Cauchy integral formula)** *If $f$ is holomorphic in a domain $\Omega$, and continuous on the closure of $\Omega$, then for each $z \in \Omega$,*

$$f(z) \;=\; \frac{1}{2\pi i}\int_{\partial\Omega}\frac{f(\zeta)}{\zeta - z}\,d\zeta. \tag{1.2.1}$$

*More generally, each derivative can be written as an integral:*

$$f^{(n)}(z) = \frac{n!}{2\pi i}\int_{\partial\Omega}\frac{f(\zeta)\,d\zeta}{(\zeta-z)^{n+1}}. \tag{1.2.2}$$

Thus a holomorphic function is infinitely differentiable. Moreover, if

$$|z-z_0| < r = \inf_{\zeta\in\partial\Omega}|\zeta-z_0|,$$

then the expansion

$$\frac{1}{\zeta-z} = \frac{1}{(\zeta-z_0)\cdot\left[1-\dfrac{z-z_0}{\zeta-z_0}\right]} = \sum_{n=0}^{\infty}\frac{(z-z_0)^n}{(\zeta-z_0)^{n+1}}$$

converges uniformly for $\zeta\in\partial\Omega$. This gives

**Theorem 1.2.2. (Taylor expansion)** *If $f$ is holomorphic in a disk $D_r(z_0)$, then $f$ has a convergent Taylor expansion*

$$f(z) = \sum_{n=0}^{\infty}a_n(z-z_0)^n, \quad |z-z_0|<r; \qquad a_n = \frac{f^{(n)}(z_0)}{n!}. \tag{1.2.3}$$

Other easy consequences of the Cauchy integral formula are various *mean value* and *maximum* principles. For example, if $f$ is holomorphic in a domain that includes the closure of a disk $D_r(z)$, then a change of variables

$$\zeta = z+re^{i\theta}$$

gives

$$f(z) = \frac{1}{2\pi i}\int_{|\zeta-z|=r}\frac{f(\zeta)}{\zeta-z}\,d\zeta = \frac{1}{2\pi}\int_0^{2\pi}f(z+re^{i\theta})\,d\theta. \tag{1.2.4}$$

One can also take the real or imaginary part of this formula.

**Theorem 1.2.3. (Mean value property)** *If $f$ is holomorphic in a domain $\Omega$, then the value of $f$ at each point $z_0\in\Omega$ is the mean of the values on any circle $\{z : |z-z_0|=r\}$ that is small enough so that $D_r(z_0)$ is contained in $\Omega$. The real and imaginary parts of $f$ have the same property.*

It is an easy consequence of Theorem 1.2.3 that the maximum value of the modulus $|f(z)|$, or of the real or imaginary parts of $f$, occurs at the boundary of $\Omega$. A closer examination of (1.2.4), taking into account the Taylor expansion, shows that no such maximum value can occur at an interior point of $\Omega$ unless $f$ is constant near the point.

**Theorem 1.2.4. (Maximum modulus principle)** *If $f$ is holomorphic in $\Omega$ and continuous on the closure of $\Omega$, then the maximum value of the modulus $|f(z)|$ is attained on the boundary. The same is true for the real and imaginary parts of $f$.*

**Theorem 1.2.5. (Strong maximum modulus principle)** *If $\Omega$ is connected and the maximum modulus is attained at a point of $\Omega$ itself, then $f$ is constant. The same is true for the real and imaginary parts of $f$.*

We note here another frequently used consequence of the Cauchy integral formula.

**Proposition 1.2.6.** *Suppose that $\{f_n\}_{n=1}^{\infty}$ is a sequence of functions holomorphic in a domain $\Omega$, and suppose that the sequence converges to a function $f$, uniformly on each compact subset of $\Omega$. Then $f$ is holomorphic in $\Omega$.*

In fact if $z \in \Omega$, the convergence is uniform on a small circle $\Gamma$ that contains $z$. Therefore in the disk bounded by $\Gamma$ the limit function $f$ is given by the Cauchy integral formula, from which it follows that $f$ is holomorphic in that disk.

An *entire function* is a function $f$ that is holomorphic in the entire plane $\mathbb{C}$. For each $R > 0$ and each $z \in \mathbb{C}$, (1.2.1) and (1.2.2) give

$$f(z) = \frac{1}{2\pi i}\int_{|\zeta - z| = R} \frac{f(\zeta)}{\zeta - z}\, d\zeta$$

and, more generally,

$$f^{(n)}(z) = \frac{n!}{2\pi i}\int_{|\zeta - z| = R} \frac{f(\zeta)}{(\zeta - z)^{n+1}}\, d\zeta.$$

Since the circle of integration has length $2\pi R$ and the modulus of the denominator is $R^{n+1}$, it is easy to see that constraints on the growth of $f$ can imply vanishing of high-order derivatives.

**Theorem 1.2.7. (Liouville's theorem)** *If $f$ is entire and bounded, then $f$ is constant.*

**Theorem 1.2.8. (Extended Liouville theorem)** *If $f$ is entire and*

$$|f(z)| \le C(|z|^n + 1)$$

*for some integer $n \ge 0$, then $f$ is a polynomial of degree $\le n$.*

## 1.3 Change of contour, isolated singularities, residues

The Cauchy integral theorem is often used to justify a *change of contour* in an integration . This is particularly useful in the rest of this section. Rather than formulate a general theorem, we illustrate with an example. Suppose that the domain $\Omega$ is bounded by one large circle $\Gamma$ and two smaller, disjoint circles, $\Gamma_1$, $\Gamma_2$, that are enclosed by $\Gamma$, as in Figure 1.1 on the left. Suppose that $f$ is holomorphic in $\Omega$ and continuous on the closure. Then

$$\int_\Gamma f(z)\,dz \;=\; \int_{\Gamma_1} f(z)\,dz + \int_{\Gamma_2} f(z)\,dz, \tag{1.3.1}$$

where each circle is oriented in the positive (counterclockwise) direction.

In fact, Theorem 1.1.1 implies that the integral of $f$ over the contour on the right in Figure 1.1 is zero. In the limit, as the gap is closed, the integrals over the flat parts of the contour cancel, and we are left with (1.3.1) in the form

$$\int_\Gamma f(z)\,dz - \int_{\Gamma_1} f(z)\,dz - \int_{\Gamma_2} f(z)\,dz \;=\; 0.$$

An *isolated singularity* for a holomorphic function is a point $z_0$ such that $f$ is holomorphic in a punctured disk $\Omega = \{z : 0 < |z - z_0| < r\}$.

An isolated singularity $z_0$ is said to be a *removable singularity* if a value $f(z_0)$ can be assigned to $f$ at $z_0$ in such a way that the extended function is holomorphic in some disk $\{z : |z - z_0| < r\}$.

An isolated singularity $z_0$ is said to be a *pole* if there is some integer $n > 0$ such that

$$f(z) \;=\; \frac{a_{-n}}{(z-z_0)^n} + \frac{a_{1-n}}{(z-z_0)^{n-1}} + \cdots + a_0 + a_1(z-z_0) + \ldots \tag{1.3.2}$$

in some punctured disk $\{0 < |z - z_0| < r\}$. The expansion (1.3.2) is called the *Laurent expansion* of $f$ at $z_0$. The *order* of the pole is $n$. A *simple pole* is a pole of order 1.

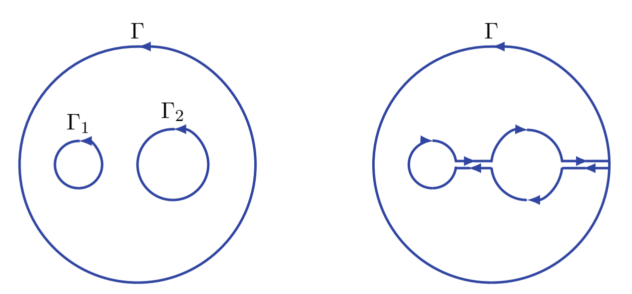



**Fig. 1.1** Change of contour in integration

Suppose that $f$ is bounded and holomorphic in the punctured disk $\{z : 0 < |z - z_0| < R\}$. By choosing a smaller radius, we may assume that $f$ is continuous up to the circle $\{z : |z - z_0| = r\}$. Let $g(z) = (z - z_0)f(z)$ and $g(z_0) = 0$, so $g(z)$ is continuous at $z_0$. Using the Cauchy integral formula for the annulus $\{z : \varepsilon < |z - z_0| < r\}$ and letting $\varepsilon \to 0$, we find that $g$ is given by the Cauchy integral formula and is therefore holomorphic near 0. If follows that the same is true for $f = g/(z - z_0)$. Thus

**Proposition 1.3.1.** *Suppose that $z_0$ is an isolated singularity of $f$ and suppose that $f(z)$ is bounded for $0 < |z - z_0| < r$. Then $z_0$ is a* removable singularity: *$f(z)$ has a limit at $z = z_0$ and extends to be holomorphic in $D_r(z_0)$.*

**Corollary 1.3.2.** *Suppose that $z_0$ is an isolated singularity of $f$. Suppose that for some integer $n$, $g(z) = (z - z_0)^n f(z)$ is bounded as $z \to z_0$, and suppose that $n$ is the least such integer. If $n$ is negative, it follows that $z_0$ is a removable singularity, at which $f$ has a zero of order $-n$. If $n$ is positive, then $f$ has a pole of order $n$ at $z_0$,*

An isolated singularity that is neither removable nor a pole is called an *essential singularity*. In this case the behavior near $z_0$ is quite different.

**Theorem 1.3.3. (Casorati–Weierstrass theorem)** *Suppose that $f$ is holomorphic in a domain $\Omega$ and has an essential singularity at $z_0 \in \Omega$. In each punctured neighborhood $D_\varepsilon = \{z : 0 < |z - z_0| < \varepsilon\}$, $f$ comes arbitrarily close to any given complex number $a$.*

*Proof:* Suppose, to the contrary, that $|f(z) - a| \geq \delta > 0$ in $D_\varepsilon$. Then $g(z) = 1/[f(z) - a]$ has an isolated singularity at $z_0$. Moreover, $g$ is bounded as $z \to z_0$, so the singularity is removable. If $g(z_0) \neq 0$, then $f$ has a removable singularity at $z_0$. If $g$ has a zero of degree $n > 0$ at $z_0$, then $f$ has a pole of order $n$ at $z_0$. □

Let us return to the Laurent expansion (1.3.2). Suppose that $f$ is holomorphic in $\{z : 0 < |z_0| < R\}$. Then $(z - z_0)^{-1-n} f(z)$ can be integrated term-by-term over the boundary of the domain $\{z : \varepsilon < |z - z_0| < r < R\}$. Taking $\varepsilon \to 0$, we find that

$$a_n = \frac{1}{2\pi i} \int_{|z-z_0|=r} \frac{f(z)\,dz}{(z - z_0)^{n+1}}. \tag{1.3.3}$$

In particular, the coefficient $a_{-1}$ is defined to be the *residue* $\mathrm{res}(f, z_0)$ of $f$ at $z_0$:

$$\mathrm{res}(f, z_0) = \frac{1}{2\pi i} \int_{|z-z_0|=r} f(z)\,dz. \tag{1.3.4}$$

A function $f$ is said to be *meromorphic* in a domain $\Omega$ if $f$ is holomorphic except at isolated points that are poles of $f$. An application of Cauchy's theorem to the domain minus sufficiently small disks centered at the poles gives the following.

**Theorem 1.3.4. (Residue theorem)** *If $f$ has finitely many poles in $\Omega$ and is continuous on the closure, then*

$$\frac{1}{2\pi i}\int_{\partial\Omega} f(\zeta)\,d\zeta \;=\; \sum_{z\in\Omega} \operatorname{res}(f,z). \tag{1.3.5}$$

The residue theorem can be used to count poles and zeros (taking into account multiplicities). In fact, suppose that near $z = z_0$, $f(z) = (z-z_0)^n g(z)$, where $n$ is an integer, $g$ is holomorphic, and $g(z_0) \neq 0$. Then

$$\frac{f'(z)}{f(z)} \;=\; \frac{n}{z-z_0} + \frac{g'(z)}{g(z)}$$

has residue $n$ at $z_0$. As a consequence:

**Theorem 1.3.5. (Counting zeros and poles)** *If $f$ is meromorphic in $\Omega$, and continuous and nowhere zero at the boundary, then*

$$\begin{aligned}&\frac{1}{2\pi i}\int_{\partial\Omega} \frac{f'(\zeta)}{f(\zeta)}\,d\zeta \\ &= \text{number of zeros of } f \text{ minus number of poles of } f \text{ in } \Omega,\end{aligned} \tag{1.3.6}$$

*where the zeros and poles are counted according to multiplicity.*

**Corollary 1.3.6.** *If $f$ is meromorphic in $\Omega$ and continuous on the boundary, then it takes each value in the complement of $f(\partial\Omega)$ the same number of times (counting multiplicity) in each connected component of this complement.*

*Proof.* If $f$ does not take the value $a$ on the boundary, then the integral

$$N(a) \;=\; \frac{1}{2\pi i}\int_{\partial\Omega} \frac{f'(\zeta)}{f(\zeta)-a}\,d\zeta$$

counts the number of times $f$ takes the value $a$ minus the number of poles. The number of poles is constant, and $N(a)$, being integer-valued and continuous with respect to $a$, is also constant on the connected component of the complement that contains $a$. □

The following are two more applications of these ideas.

**Theorem 1.3.7. (Rouché's theorem)** *Suppose that $f$ and $g$ are holomorphic in $\Omega$ and continuous on the closure. If $|f(z)-g(z)| < |f(z)|$ on the boundary $\partial\Omega$, then $f$ and $g$ have the same number of zeros in $\Omega$.*

In fact the function $f_s(z) = (1-s)f(z) + sg(x) = f(z) - s[f(z)-g(z)]$, $0 \le s \le 1$, has no zeros on $\partial\Omega$, so the number of zeros in $\Omega$ is

$$\frac{1}{2\pi i}\int_{\partial\Omega}\frac{f_s'(\zeta)}{f_s(\zeta)}\,d\zeta.$$

This is an integer-valued continuous function of $s$, so it has the same value at $s = 0$ and at $s = 1$. But $f_0 = f$, $f_1 = g$.

**Theorem 1.3.8. (Inverse function theorem)** *Suppose that $f$ is holomorphic near $z_0$ and $f'(z_0) \neq 0$. Then $f$ has an inverse that is holomorphic near $f(z_0)$.*

In fact it follows from the series expansion at $z_0$ that for small $r > 0$, $f(z) \neq f(z_0)$ if $z$ is inside or on the curve $\Gamma = \{z : |z - z_0| = r\}$. Therefore if $a$ is close enough to $f(z_0)$, the integral

$$\frac{1}{2\pi i}\int_{\Gamma}\frac{\zeta f'(\zeta)}{f(\zeta)-a}\,d\zeta$$

is the unique value of $z$ inside the curve such that $f(z) = a$. This expression is a holomorphic function of $a$.

## 1.4 The logarithm and powers

In view of (1.1.1), the complex *logarithm* $\log z$, $z \neq 0$, is defined by

$$\log z \;=\; \log(|z|e^{i\arg z}) \;=\; \log|z| + i\arg z. \tag{1.4.1}$$

Here $\log|z|$ denotes the usual choice for positive argument; thus $\log|z|$ is real. Of course $\arg z$ is defined only up to addition of an integer multiple of $2\pi$. By a *branch* of the logarithm in a connected domain $\Omega$, we mean a choice that is holomorphic throughout $\Omega$. (Such a choice may not be possible, e.g. in a deleted neighborhood of the origin $\{z : 0 < |z| < r\}$.) A branch is called the *principal branch* if $\Omega \cap \mathbb{R}$ is not empty and $\log z$ is real on this intersection.

An important concept here is that of a *simply connected* domain, usually defined to be one that is connected and in which each closed curve can be continuously shrunk to a point. An equivalent definition is that $\Omega$ is connected and given two curves $\gamma_0$ and $\gamma_1$ in $\Omega$ that join points $z$ and $w$, there is a family of curves $\gamma_t : [0,1] \to \Omega$, $0 < t < 1$, such that $\gamma_t(0) = z$, $\gamma_t(1) = w$, and the map $(s,t) \to \gamma_t(s)$ is continuous, $0 \leq s,t \leq 1$. (Showing that the two definitions are equivalent is an interesting exercise.)

Suppose that $\Omega$ is a simply connected domain. Suppose also that 0 is not in $\Omega$. Then a branch of the logarithm may be obtained by choosing $z_0 \in \Omega$, choosing $\log z_0$, and setting

$$\log z \;=\; \log z_0 + \int_{z_0}^{z}\frac{d\zeta}{\zeta}. \tag{1.4.2}$$

Because of the assumption that $\Omega$ is simply connected, the integral is independent of the path of integration from $z_0$ to $z$: see Section 1.7 for details.

Corresponding to a branch of the logarithm, and to each $\alpha \in \mathbb{C}$, there is a branch of the power $z^\alpha$:

$$z^\alpha \;=\; e^{\alpha \log z}. \tag{1.4.3}$$

This is independent of the branch of the logarithm if and only if $\alpha$ is an integer.

## 1.5 Infinite products

We will encounter a number of *infinite products*, often written in the form

$$\prod_{n=1}^{\infty}(1-a_n), \tag{1.5.1}$$

where the $a_n$ are complex numbers. The key tool to be used is the following estimate.

**Lemma 1.5.1.** *Suppose* $|z| \le 1/2$. *Then the principal branch of* $\log(1-z)$ *satisfies*

$$|\log(1-z)+z| \;\le\; |z|^2 \;\le\; \frac{|z|}{2}. \tag{1.5.2}$$

*Proof:* Integrating along the line segment from 1 to $1+z$,

$$\begin{aligned}\log(1-z) &= \int_1^{1-z}\frac{ds}{s} \;=\; -\int_0^z \frac{dt}{1-t}\\ &= -\int_0^z (1+t+\dots)\,dt \;=\; -z-\frac{z^2}{2}-\dots,\end{aligned}$$

so

$$|\log(1-z)+z| \;\le\; \frac{|z|^2}{2}(1+|z|+|z|^2+\dots) \;=\; \frac{|z|^2}{2}\cdot\frac{1}{1-|z|} \;\le |z|^2. \qquad \square$$

The (formal) product (1.5.1) is said to converge if

$$\lim_{M,N\to\infty}\prod_{n=M}^{N}(1-a_n) \;=\; 1. \tag{1.5.3}$$

This implies that the partial products $\prod_M^\infty(1-a_n)$ have a non-zero limit, as soon as $M$ is large enough that $n \ge M$ implies $1-a_n \ne 0$. In particular, a *necessary* condition for convergence is that $1-a_n \to 1$, i.e. $a_n \to 0$. Suppose that $|a_n| \le 1/2$ for $n \ge M$. Then, taking the principal branch of the logarithm

$$\log\left|\prod_M^N(1-a_n)\right| \;=\; \sum_{n=M}^{N}|\log(1-a_n)|.$$

The product is said to be *absolutely convergent* if

$$\prod_{n=1}^{\infty}(1+|a_n|)$$

converges. Absolute convergence implies convergence. It follows from (1.5.2) that if $|a_n| \le 1/2$, then

$$\frac{|a_n|}{2} \le |\log(1+|a_n|) \le \frac{3|a_n|}{2}.$$

Therefore the product converges absolutely if and only if $\sum_{n=1}^{\infty}|a_n| < \infty$.

We mention here the *Weierstrass product theorem*: if $\{a_n\}$ is a sequence of points in the plane such that $|a_n|$ tends to $\infty$, then there is an entire function whose zeros, counting multiplicity, are the points $a_n$. For the proof, see Chapter 8. Here is a proof for a case that is encountered several times in other chapters.

**Theorem 1.5.2.** *Suppose that $\{a_n\}$ is a sequence of non-zero numbers such that the sum $\sum_{n=1}^{\infty}|a_n|^{-2}$ is finite. Then the product*

$$f(z) = \prod_{n=1}^{\infty}\left(1-\frac{z}{a_n}\right)e^{z/a_n} \tag{1.5.4}$$

*defines an entire function whose zeros are the $a_n$.*

*Proof:* For a given $z$, $|z/a_n|$ will be less than $1/2$ for $n \ge N(z)$. The logarithm of the $n$-th factor is

$$\log\left(1-\frac{z}{a_n}\right)+\frac{z}{a_n} \tag{1.5.5}$$

and by (1.5.2), for fixed $z$ (1.5.5) has modulus $\le |z|^2/|a_n|^2$ for large $n$. Thus the product converges uniformly on bounded sets. This implies that the function $f$ is entire. Moreover, the product vanishes only where some factor vanishes. $\square$

## 1.6 Reflection principles

**Theorem 1.6.1.** *Suppose that $\Omega$ is a domain that is symmetric under reflection about the real axis: $\overline{\Omega} = \Omega$, and the intersection $I = \Omega \cap \mathbb{R}$ is not empty. Suppose also that $f$ is holomorphic on the intersection of $\Omega$ with the upper half-plane $\mathbb{C}_+ = \{z : \operatorname{Im} z > 0\}$, continuous up to $I$, and real on $I$. Then $f$ has a holomorphic extension to the remainder of $\Omega$, with*

$$f(\bar{z}) = \overline{f(z)}, \quad z \in \Omega_+. \tag{1.6.1}$$

*Proof:* The prescription (1.6.1) defines $f$ so as to be holomorphic in $\Omega \cap \mathbb{C}_-$, and continuous in all of $\Omega$. We need to show that $f$ is holomorphic near $I$. Consider a complex neighborhood $D_r(x_0)$ of a point $x_0 \in I$, whose closure is contained in $\Omega$.

Let

$$\Delta_{\pm} = D_r(x_0) \cap \{z : \pm \mathrm{Im}\, z > 0\}, \tag{1.6.2}$$

and

$$g(z) = \frac{1}{2\pi i} \int_{|\zeta - x_0| = r} \frac{f(\zeta)\, d\zeta}{\zeta - z}, \qquad |z - x_0| < r.$$

This function is holomorphic in $D_r(x_0)$. For $z \in \Delta_+$ the lower semicircle of the contour can be moved to the $x$-axis, showing that $g = f$ on $\Delta_+$. Similarly, $g = f$ on $\Omega_-$ if we use (1.6.1) to define $f$ on $\Delta_-$. It follows that (1.6.1) extends $f$ holomorphically across $I$. □

**Theorem 1.6.2.** *Suppose that $\Omega$ and $I$ are as in the previous theorem. Suppose that $f$ is holomorphic in $\Omega \cap \mathbb{C}_+$, nowhere zero, and continuous up to $I$. Suppose also that $|f(x)| = 1$ for $x \in I$. Then $f$ has a holomorphic extension to the remainder of $\Omega$, with*

$$f(\bar{z}) = 1/\overline{f(z)}. \tag{1.6.3}$$

*Proof:* As in the previous proof, it is sufficient to work in a small disk $D_r(x_0)$. For small $r$ a branch $g$ of $\log f$ can be chosen in $\Delta_+$. By the assumption on $|f|$, the limit of $ig$ is real on $D_r(x_0) \cap \mathbb{R}$. Therefore $ig$ can be continued to all of $D_r(x_0)$. The continuation of $g$, given by (1.6.1) for $ig$, exponentiates to the continuation of $f$ given by (1.6.3). □

The assumption in Theorem 1.6.2 that $f$ is continuous up to $I$ and $|f(z)| = 1$ on $I$ can be replaced by the weaker assumption that $|f(z)| \to 1$ as $z$ approaches I. This is the *Schwarz reflection theorem*; see Chapter 4. It plays a key role in connection with conformal mapping in Chapters 5, 6, and 17.

## 1.7 Analytic continuation

There are two situations that give rise to the consideration of analytic continuation. An example of one such situation is the function $f$ defined by the series

$$f(z) = 1 + z + z^2 + z^3 + \cdots + z^n + \ldots \,. \tag{1.7.1}$$

The series converges if and only if $|z| < 1$. On the other hand, the sum is $1/(1-z)$, which is holomorphic in the complement of the point $z = 1$. It is natural to consider $1/(1-z)$ as a *continuation* of $f$: the extension of $f$ to a function holomorphic on a larger domain. A natural question: is such an extension unique?

An example of a second such situation is the logarithm. Starting with the usual choice in a neighborhood of $z = 2$, and following along a curve that circles the origin in the positive direction, one comes back not to $\log 2$ but to $\log 2 + 2\pi i$—but it is natural to think of this "branch" as an analytic continuation of the original. For a visualization see Figure 1.2.

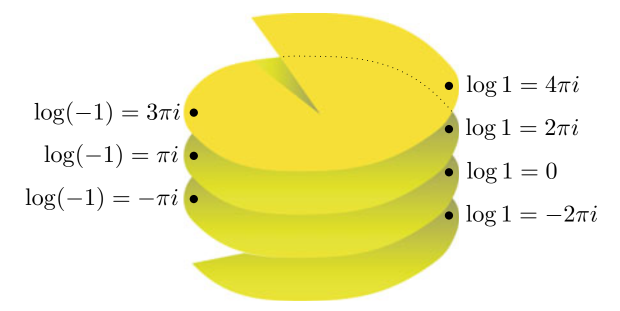



**Fig. 1.2** Analytic continuation of the logarithm

In general, suppose that $f_0$ is holomorphic in an open disk $D_0$ centered at $z_0$, suppose that $\gamma : [0,1] \to \mathbb{C}$ is a curve with $\gamma(0) = z_0$, and suppose that $D_0$ does not contain $\gamma$ (we are systematically conflating $\gamma$ as a mapping and $\gamma$ as a set of points, i.e. the image of the mapping). It may still be the case that we can find successive points $z_j = \gamma(t_j)$ along the curve and functions $f_j$ holomorphic in disks $D_j$ centered at $z_j$ such that $D_j \cap D_{j+1} \neq \emptyset$, $f_j = f_{j+1}$ on $D_j \cap D_{j+1}$, and the union of the $D_j$ covers $\gamma$. The result is a function $f$, holomorphic in a neighborhood of the curve $\gamma$, that agrees with $f_0$ near $z_0$. The function $f$ is said to be a *continuation of $f_0$ along the curve $\gamma$*.

**Proposition 1.7.1.** (Uniqueness of analytic continuation) *If two functions that are holomorphic in a connected domain $\Omega$ agree on a non-empty open subset of $\Omega$, then they agree on all of $\Omega$.*

*Proof:* It suffices to prove that if $f$ is holomorphic in $\Omega$ and vanishes near a point $z_0 \in \Omega$, then $f$ is identically zero. Let $z$ be another point of $\Omega$ and let $\gamma : [0,1] \to \Omega$ be a smooth curve with $\gamma(0) = z_0$ and $\gamma(1) = z$. If $f(\gamma(s)) = 0$ for $0 \leq s \leq t$, then it follows that each derivative of $f$ vanishes at $z = \gamma(t)$. Thus the Taylor expansion of $f$ vanishes at $\gamma(t)$, so $f$ vanishes in a neigborhood of $\gamma(t)$. It follows from this argument that $f$ vanishes along the entire curve, so $f(z) = 0$. □

Recall from Section 1.4: a connected domain $\Omega \subset \mathbb{C}$ is said to be simply connected if each closed curve in $\Omega$ can be deformed continuously to a point (a constant curve). For example, the plane $\mathbb{C}$ is simply connected, but the complement of any non-empty bounded subset $A$ is not. As noted above, an equivalent definition is that any two curves from a point $z_0$ to a point $z_1$ can be deformed continuously from one to the other.

**Theorem 1.7.2. (Monodromy theorem)** *Suppose that the domain $\Omega$ is simply connected. Suppose that $f_0$ is holomorphic in a domain $\Omega_0 \subset \Omega$, and suppose that $f_0$ can be continued along each curve in $\Omega$. Then $f_0$ has a unique holomorphic extension to all of $\Omega$.*

*Proof:* Take $z_0 \in \Omega_0$. It is enough to show that the continuation of $f_0$ along a curve $\gamma\colon [0,1] \to \Omega$ that starts at $z_0$ leads to a value $f(\gamma(1))$ that depends only on $z_1 = \gamma(1)$, not on the particular curve $\gamma$. Suppose that $\gamma_0$ and $\gamma_1$ are two such curves from $z_0$ to $z_1$. Then there is a family of curves $\gamma_t$ from $z_0$ to $z_1$, $0 < t < 1$, that interpolates continuously from $\gamma_0$ to $\gamma_1$.

Suppose that $f_0$ is continued along each curve $\gamma_t$. Let $T$ be the supremum of those $t$ such that $\gamma_t(z_1) = \gamma_0(z_1)$ It follows from Proposition 1.7.1 that $T$ is positive. It follows from continuity that $\gamma_T(z_1) = \gamma_0(z_1)$. Then $T = 1$, since, otherwise, Proposition 1.7.1 implies that equality at $z_1$ can be extended past $t = T$. □

## 1.8 The Stieltjes integral

The basic idea of the Stieltjes integral is to weight an interval $I = (a,b)$ not by its length $b-a$ but by $g(b)-g(a)$, where $g$ is some non-decreasing real function. Then ordinary Riemann sums are replaced by sums

$$\sum_{j=1}^{n} f(x_j)\,[g(x_j)-g(x_{j-1})], \quad a = x_0 < x_1 < \cdots < x_n = b,$$

and the limit as the intervals shrink is denoted

$$\int_a^b f(x)\,dg(x).$$

With some attention to behavior at the endpoints, one can integrate by parts, assuming that $f$ is continuously differentiable, leading to

$$\int_a^b f(x)\,dg(x) \;=\; f(x)\,g(x)\Big|_a^b - \int_a^b g(x)\,f'(x)\,dx. \tag{1.8.1}$$

We will not need the general theory, only the special case when $g$ is piecewise constant. Say $g : [0,\infty) \to [0,\infty)$,

$$g(x) = \sum_{x_k \le x} c_k, \quad 0 < x_1 < x_2 < \cdots < x_n < \ldots.$$

We take the integral to be continuous from the right. Then it is easily checked that

$$\int_0^x f(x)\,dg(x) \;=\; \lim_{\varepsilon\to 0+}\int_0^{x+\varepsilon} f(x)\,dg(x) \;=\; \sum_{x_k\le x} c_k\,f(x_k). \tag{1.8.2}$$

In this case one can justify (1.8.1) by simple bookkeeping. If $x_n < x < x_{n+1}$ then

$$\begin{aligned}
\int_0^x g(t)f'(t)\,dt &= \sum_{j=1}^{n-1}\int_{x_j^+}^{x_{j+1}^-} g(t)f'(t)\,dt + \int_{x_n^+}^{x} g(t)f'(t)\,dt \\
&= \sum_{j=1}^{n-1} g(x_j)[f(x_{j+1})-f(x_j)] + g(x_n)[f(x)-f(x_n)] \\
&= -g(x_1)f(x_1) - \left\{\sum_{j=1}^{n-1}[g(x_{j+1})-g(x_j)]f(x_{j+1})\right\} + g(x)f(x) \\
&= -\sum_{j=1}^{n} c_j f(x_j) + g(x)f(x). \qquad (1.8.3)
\end{aligned}$$

This is the discrete version of the integration-by-parts formula (1.8.1).

Summarizing,

**Proposition 1.8.1.** *Suppose that $f:[0,\infty)\to\mathbb{C}$ is continuously differentiable and*

$$g(x) = \sum_{x_k\le x} c_k,$$

*where $c_k > 0$ and $0 < x_1 < x_2 < x_3 < \cdots < x_n < \ldots$ . Then*

$$\int_0^{x+} f(t)\,dg(t) \;=\; \sum_{x_k\le x} c_k f(x_k) \;=\; f(x)\,g(x) - \int_0^{x+} g(t)f'(t)\,dt, \qquad (1.8.4)$$

## 1.9 Hilbert spaces

For some applications we need the basics of Hilbert space theory. The starting point is an *inner product space*. This is a complex vector space $H$, equipped with an *inner product* $(u,w)$, defined for each pair $u$, $w$ in $H$ and having the properties

$$(a_1u_1+a_2u_2,w) \;=\; a_1(u_1,w)+a_2(u_2,w), \quad a_j\in\mathbb{C},\;\; u_j,w\in H; \qquad (1.9.1)$$

$$(u,w) \;=\; \overline{(w,u)}, \qquad u,w\in H; \qquad (1.9.2)$$

$$(u,u) \;>\; 0 \quad \text{if}\;\; u\in H \;\;\text{and}\;\; u\ne 0. \qquad (1.9.3)$$

Let

$$||u|| = \sqrt{(u,u)}.$$

A basic property is the *Cauchy–Schwarz inequality*

$$|(u,w)| \le ||u||\,||w||. \qquad (1.9.4)$$

The proof can be reduced to the case $||u|| = ||w|| = 1$. Then for each $a\in\mathbb{C}$ with $|a| = 1$,

$$\begin{aligned} 0 \le ||u - aw||^2 &= (u - aw, u - aw) \\ &= ||u||^2 - (u, aw) - (aw, u) + ||aw||^2 \\ &= 2 - 2\,\text{Re}\,\{\bar{a}(u, w)\}. \end{aligned}$$

We may choose $a$ with $|a| = 1$ in such a way that $\text{Re}\,\{\bar{a}(u, w)\} = |(u, w)|$.

The Cauchy–Schwarz inequality implies the triangle inequality

$$||u + w|| \le ||u|| + ||w||.$$

This and the positivity property (1.9.3) imply that $d(u, w) = ||u - w||$ is a metric. The space $H$ is said to be a *Hilbert space* if $H$ is complete with respect to this metric.

First example: the space $l^2(\mathbb{Z})$ of two-sided complex sequences $\mathbf{x} = (x_n)_{-\infty}^{\infty}$ such that

$$\sum_{n=-\infty}^{\infty} |x_n|^2 < \infty,$$

with inner product

$$(\mathbf{x}, \mathbf{y}) = \sum_{n=-\infty}^{\infty} x_n \bar{y}_n.$$

The Cauchy–Schwarz inequality, applied to partial sums, implies that the inner product is well-defined. This space is easily shown to be complete.

Second example: the space of continuous functions $u : \mathbb{R} \to \mathbb{C}$ that are periodic, with period $2\pi$:

$$u(x + 2\pi) = u(x), \qquad (u, w) = \frac{1}{2\pi} \int_{-\pi}^{\pi} u(x)\overline{w(x)}\, dx.$$

The completion of this inner product space with respect to the associated metric is $L^2_{\text{per}}(\mathbb{R})$; it can be identified with the corresponding space for the interval $[0, 2\pi]$,

$$L^2([0, 2\pi]).$$

Two elements $u$, $w$ of an inner product space are said to be *orthogonal*, written $u \perp w$, if $(u, w) = 0$. Note that

$$u \perp w \;\Rightarrow\; ||u + w||^2 = ||u||^2 + ||w||^2.$$

An *orthonormal set* in an inner product space $H$ is a subset consisting of elements $\{\varphi_j\}$ such that

$$(\varphi_j, \varphi_k) = \begin{cases} 1 \text{ if } & j = k; \\ 0 \text{ if } & j \ne k. \end{cases}$$

For our purposes the index set $\{j\}$ here is finite or countable. Let us suppose that it is the integers $\mathbb{Z}$. If $\{\varphi_n\}$ is an orthonormal set in $H$, let

$$u_n = \sum_{|j|\le n} (u,\varphi_j)\,\varphi_j.$$

An easy calculation shows that $u_n$ and $u-u_n$ are orthogonal, so

$$\sum_{|j|\le n} |(u,\varphi_j)|^2 = ||u_n||^2 = ||u||^2 - ||u-u_n||^2.$$

This implies *Bessel's inequality*:

$$\sum_{|j|\le n} |(u,\varphi_j)|^2 \le ||u||^2, \tag{1.9.5}$$

and also *Bessel's equality*:

$$||u-u_n|| \to 0 \Leftrightarrow \sum_{j=-\infty}^{\infty} |(u,\varphi_j)|^2 = ||u||^2. \tag{1.9.6}$$

The orthonormal set $\{\varphi_j\}$ is said to be *complete*, or *an orthonormal basis* if $u_n$ converges to $u$ for every $u \in H$. Note that $u_n$ is the element closest to $u$ in the subspace $H_n$ spanned by $\{\varphi_j\}_{-n}^{n}$. In fact if $w$ belongs to $H_n$, then

$$||u-(u_n+w)||^2 = ||u-u_n||^2 + ||w||^2$$

is minimal when $w=0$.

## 1.10 $L^p$ spaces

We consider two examples that illustrate the basic principles. The first is the space $l^p(\mathbb{Z})$ of two-sided complex sequences $\mathbf{x} = (x_n)_{-\infty}^{\infty}$ such that

$$\sum_{n=-\infty}^{\infty} |x_n|^p < \infty.$$

Here it is assumed that $1 \le p < \infty$. For $p=\infty$ the condition is replaced by the condition that $\sup|x_n| < \infty$. We define

$$||\mathbf{x}||_p = \left[\sum_{n=-\infty}^{\infty} |x_n|^p\right]^{1/p}, \qquad 1 \le p < \infty \tag{1.10.1}$$

and

$$||\mathbf{x}||_\infty = \sup|x_n|.$$

It is easily seen that $||\mathbf{x}||_p$ defines a norm when $p=1$ or $p=\infty$. For intermediate values we use *Hölder's inequality*. Again, let

$$(\mathbf{x}, \mathbf{y}) = \sum_{n=-\infty}^{\infty} x_n \bar{y}_n.$$

Then Hölder's inequality is

$$|(\mathbf{x}, \mathbf{y})| \leq ||\mathbf{x}||_p \, ||\mathbf{y}||_q \quad \text{if} \quad 1 \leq p \leq \infty \quad \text{and} \quad \frac{1}{p} + \frac{1}{q} = 1. \tag{1.10.2}$$

This is a partial generalization of the Cauchy–Schwarz inequality—the case $p = 2$ here—and it is obvious for $p = 1$ or $p = \infty$. Otherwise we reduce to the case $||\mathbf{x}||_p = ||\mathbf{y}||_q = 1$ and start from the elementary inequality

$$|ab| \leq \frac{|a|^p}{p} + \frac{|b|^q}{q} \quad \text{if} \quad p, q > 0, \text{ and } \frac{1}{p} + \frac{1}{q} = 1.$$

Plugging in $x_n$ and $\bar{y}_n$ for $a$ and $b$ and summing gives (1.10.2). The next step is to note that Hölder's inequality is best possible: if $1 \leq p \leq \infty$, then

$$||\mathbf{x}||_p = \sup_{||\mathbf{y}||_q = 1} |(\mathbf{x}, \mathbf{y})|, \qquad \frac{1}{p} + \frac{1}{q} = 1. \tag{1.10.3}$$

Again this is easily seen in the extreme cases $p = 1$ and $p = \infty$. If $1 < p < \infty$ we assume again that $||\mathbf{x}||_p = 1$. Set $y_n = 0$ if $x_n = 0$, and otherwise let

$$y_n = |x_n|^{p-1} \frac{x_n}{|x_n|}.$$

Then $(\mathbf{x}, \mathbf{y}) = ||\mathbf{x}||_p^p = ||\mathbf{y}||_q^q = 1$. The triangle inequality follows:

$$||\mathbf{x} + \mathbf{y}||_p = \sup_{||\mathbf{z}||_q = 1} |(\mathbf{x} + \mathbf{y}, \mathbf{z})| \leq ||\mathbf{x}||_p + ||\mathbf{y}||_p.$$

Thus (1.10.3) is a norm and defines a metric. The spaces here are complete.

The second example starts with the space of functions $u : \mathbb{R} \to \mathbb{C}$ that have period $2\pi$, with

$$||u||_p = \left[ \frac{1}{2\pi} \int_{-\pi}^{\pi} |u(x)|^p \, dx \right]^{1/p}, \tag{1.10.4}$$

where $1 \leq p < \infty$. The discussion of $l^p(\mathbb{Z})$ can be adapted to show that Hölder's inequality holds in this case:

$$\left| \frac{1}{2\pi} \int_{-\pi}^{\pi} u(x) \overline{w(x)} \, dx \right| \leq ||u||_p ||w||_q, \quad 1 < p < \infty, \quad \frac{1}{p} + \frac{1}{q} = 1. \tag{1.10.5}$$

Again, the triangle inequality is a consequence, so (1.10.4) is a norm and defines a metric. The completion of the space of periodic continuous functions with respect to this metric is the space $L^p_{\text{per}}(\mathbb{R})$, or equivalently $L^p([0, 2\pi])$.

## Remarks and further reading

Most undergraduate textbooks on complex analysis cover the basic complex analysis in this chapter. Graduate texts on real analysis or functional analysis generally have reasonably thorough treatments of Hilbert space theory and $L^p$ spaces.

Three classic complex analysis textbooks cover not only the material in this chapter but several of the topics in subsequent chapters: Ahlfors [5], Hille [64], and Titchmarsh [134].

The special issue of the Journal *Primus*, vol. 27, issue 8–9 (2017) on "Revitalizing Complex Analysis" contains a number of papers that explore topics in this chapter and their applications.

# Chapter 2
# Linear Fractional Transformations



In this chapter we introduce the Riemann sphere $\mathbb{S}$ and the meromorphic functions that map $\mathbb{S}$ bijectively to itself. These transformations play many roles in complex analysis, and in its applications to geometry and algebra. Of particular importance are the transformations that map the upper half plane to itself, and the transformations that map the unit disk to itself.

## 2.1 The Riemann sphere

The complex plane can be completed by adding the point at infinity, denoted $\infty$. By definition, a neighborhood of $\infty$ is a set that contains $\infty$ and a set $\{z : |z| > R\}$, for some $R \geq 0$. Topologically, the resulting surface is a 2-sphere. The standard pictorial representation is obtained by considering $\mathbb{C}$ as the $x, y$ plane of the three-dimensional space

$$\mathbb{R}^3 = \mathbb{C} \times \mathbb{R} = \{(w,t) : w \in \mathbb{C}, t \in \mathbb{R}\},$$

and relating it to the 2-sphere of radius 1 centered at the origin:

$$S = \{(w,t) : |w|^2 + t^2 = 1\}.$$

The relation is *stereographic projection*: a point $\omega = (w,t)$ on $S$ is mapped to a point $z = \pi(\omega)$ in $\mathbb{C}$ by following the line from the north pole $N = (0,1) \in S$ through $(w,t)$ to its intersection $(z,0)$ with $\mathbb{C} \times \{0\}$; see Figure 2.1.

The line determined by $(0,1)$ and $\omega = (w,t)$ is the set of points

$$(1-\lambda)(0,1) + \lambda(w,t), \quad \lambda \in \mathbb{R}.$$

Thus for $t \neq 1$,

$$\pi(w,t) = \frac{w}{1-t}.$$

It follows that



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_2

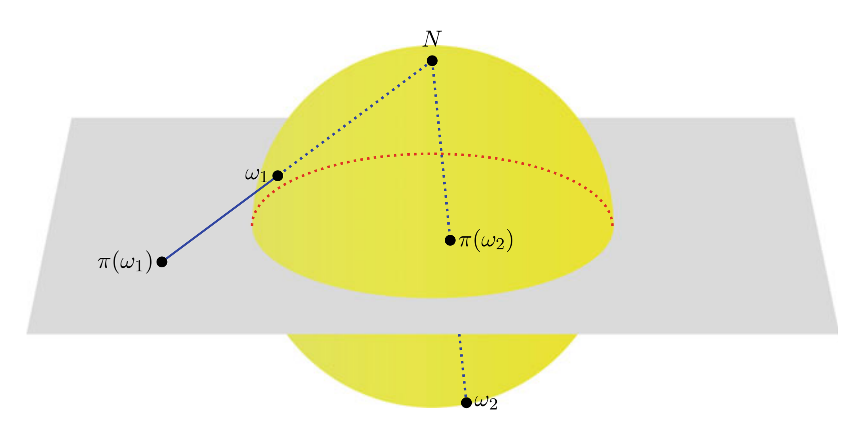



**Fig. 2.1** Stereographic projection

$$|\pi(w,t)|^2 = \frac{|w|^2}{(1-t)^2} = \frac{1-t^2}{(1-t)^2} = \frac{1+t}{1-t}.$$

This can be solved for $t$ as a function of $z = \pi(w,t)$, showing that the inverse map $\pi^{-1}$ from $\mathbb{C}$ to $S$ is given by

$$\pi^{-1}(z) = ((1-t)z, t) = \left(\frac{2z}{|z|^2+1}, \frac{|z|^2-1}{|z|^2+1}\right).$$

As $\omega \in S$ approaches the north pole, $\pi(\omega)$ approaches $\infty$, so we let $\pi(0,1) = \infty$.

Some properties of the projection $\pi$ are developed in the exercises. One property is that the image of a circle in $S$ is either a circle or a line in the plane, and conversely; see Exercise 3. Another is that if we define a distance function in the plane by using the euclidean distance of the pullback to $\mathbb{C}$,

$$d(z_1,z_2) = c\,|\pi^{-1}(z_1) - \pi^{-1}(z_2)|, \tag{2.1.1}$$

for some choice of a scaling constant $c > 0$, then

$$d(z_1,z_2) = c \cdot \frac{2|z_1 - z_2|}{\sqrt{1+|z_1|^2}\sqrt{1+|z_2|^2}}; \tag{2.1.2}$$

see Exercise 4. We choose $c = 1/2$ so that the distance function

$$d(z_1,z_2) = \frac{|z_1 - z_2|}{\sqrt{1+|z_1|^2}\sqrt{1+|z_2|^2}} \tag{2.1.3}$$

is close to the euclidean distance when the $z_j$ are close to $z = 0$.

Taking the limit as $z_2 \to \infty$ gives

$$d(z,\infty) = \frac{1}{\sqrt{1+|z|^2}}.$$

From now on we use $\mathbb{S}$ to denote the *Riemann sphere*: the set $\mathbb{C} \cup \{\infty\}$. We can give $\mathbb{S}$ a complex structure by specifying what it means for a function to be holomorphic in a neighborhood of $\infty$. We say that $f$, defined in a punctured neighborhood of $\infty$, $\{z : |z| > R\}$, is holomorphic at $\infty$ if the function $g(z) = f(1/z)$, which is defined in the domain $\{z : 0 < |z| < 1/R\}$, is holomorphic and has a removable singularity at $z = 0$. Similarly, we can define the notion of a pole or an essential singularity at $\infty$. If a function $f$ has a pole at a point $z$ of $\mathbb{S}$, we define $f(z) = \infty$. In particular, a function that is meromorphic on (all of) $\mathbb{S}$ can be considered as being everywhere defined as a map from $\mathbb{S}$ to itself.

A natural question is: what meromorphic functions on $\mathbb{S}$ are bijective (one-to-one and onto)?

**Proposition 2.1.1.** *A meromorphic function $f$ on $\mathbb{S}$ is bijective if and only if $f$ can be expressed in the form*

$$f(z) \;=\; \frac{az+b}{cz+d}, \tag{2.1.4}$$

*where $a, b, c, d$ are complex constants and $ad - bc \neq 0$.*

*Proof:* Suppose first that $f$ is bijective. By assumption, $f$ has at exactly one pole, say at $z_0$. A consequence of Corollary 1.3.6 is that if this pole has order $k$, then, near $z_0$, $f$ takes each sufficiently large value exactly $k$ times. Therefore $k = 1$: the pole is simple.

Suppose first that the pole is located at $z_0 = \infty$, and that the residue is $a \neq 0$. Then $g(z) = f(z) - az$ is an entire function that is bounded at $\infty$ and so, by Liouville's theorem, is constant:

$$f(z) \;=\; az + b, \qquad a \neq 0. \tag{2.1.5}$$

Otherwise, $f$ has a simple pole at some finite point $z = d$ with a residue $c \neq 0$. Then $f(z) - c/(z - d)$ is bounded and entire, hence constant:

$$f(z) \;=\; a + \frac{c}{z-d}, \qquad c \neq 0. \tag{2.1.6}$$

Each of the expressions (2.1.5) and (2.1.6) can easily be written in the form (2.1.4) (with different constants).

Conversely, suppose that $f$ has the form (2.1.4). Then it is easily checked that the equation $f(z) = w$ has a unique solution for each $w \in \mathbb{S}$ if and only if $ad - bc \neq 0$.

□

Note that the expression (2.1.4) for a given map $f$ is not unique: numerator and denominator can be multiplied by the same non-zero constant. It follows that we may always assume that $ad - bc = 1$.

Transformations of the form (2.1.4) are called *Möbius transformations* or *linear fractional transformations*.

It is natural to associate to the linear fractional transformation (2.1.4) a matrix $A$ and write

$$f_A(z) = \frac{az+b}{cz+d}, \qquad A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}. \tag{2.1.7}$$

The condition $ad - bc \neq 0$ is precisely the condition that $A$ be invertible.

**Proposition 2.1.2.** *The composition of linear transformations $f_A$ and $f_B$ is given by*

$$f_A \circ f_B = f_{AB}. \tag{2.1.8}$$

The proof is a simple calculation. For a more conceptual approach, we make a brief excursion into *complex projective space*. Start with the standard two-dimensional complex vector space $\mathbb{C}^2$, with the elements written as column vectors. Consider the collection of *complex lines* through the origin, i.e. sets of the form

$$L_{z,w} = \left\{ \lambda \begin{bmatrix} z \\ w \end{bmatrix} : \lambda \in \mathbb{C} \right\},$$

for some pair $(z,w) \neq (0,0)$. We may identify such a line with the equivalence class of an element of $\mathbb{C}^2 \setminus \{0\}$ where

$$\begin{bmatrix} z' \\ w' \end{bmatrix} \sim \begin{bmatrix} z \\ w \end{bmatrix} \quad \text{if} \quad \begin{bmatrix} z' \\ w' \end{bmatrix} = \mu \begin{bmatrix} z \\ w \end{bmatrix}$$

for some $\mu \in \mathbb{C}$, $\mu \neq 0$. This expresses the fact that these two points of $\mathbb{C}^2$ lie on the same line through the origin. Then

$$\begin{bmatrix} z \\ w \end{bmatrix} \sim \begin{bmatrix} z/w \\ 1 \end{bmatrix} \text{ if } w \neq 0; \qquad \begin{bmatrix} z \\ 0 \end{bmatrix} \sim \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ if } z \neq 0.$$

Identifying $[z,1]^t$ with $z \in \mathbb{C}$ and $[1,0]^t$ with $\infty$ identifies this space of complex lines with the Riemann sphere $\mathbb{S}$.

If $A : \mathbb{C}^2 \to \mathbb{C}^2$ is a linear transformation, then $A$ commutes with multiplication by $\lambda \in \mathbb{C}$, so $A$ takes equivalent elements to equivalent elements. In other words, $A$ induces a mapping of $\mathbb{S}$ to itself. In terms of canonical representatives of the form $[z,1]^t$, if the matrix of $A$ is

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

then, when $cz + d \neq 0$, we have

$$A \begin{bmatrix} z \\ 1 \end{bmatrix} = \begin{bmatrix} az+b \\ cz+d \end{bmatrix} \sim \begin{bmatrix} (az+b)/(cz+d) \\ 1 \end{bmatrix} = \begin{bmatrix} f_A(z) \\ 1 \end{bmatrix}.$$

Thus linear fractional transformations are precisely the transformations of the Riemann sphere $\mathbb{S}$ that are induced by invertible linear transformations of $\mathbb{C}^2$, when $\mathbb{S}$ is identified with the space of complex lines through the origin in $\mathbb{C}^2$. A consequence of this correspondence is a conceptual proof (or a framework) for the identity $f_A \circ f_B = f_{AB}$.

An important property of linear fractional transformations is the following. The proof is left as Exercise 12.

**Proposition 2.1.3.** *Given a triple* $(z_1, z_2, z_3)$ *of distinct points in* $\mathbb{S}$*, there is a unique linear fractional transformation* $f$ *such that*

$$f(z_1) = 0, \quad f(z_2) = 1, \quad f(z_3) = \infty. \tag{2.1.9}$$

**Corollary 2.1.4.** *Any given triple* $(z_1, z_2, z_3)$ *of distinct points in* $\mathbb{S}$ *can be taken to any other such triple by a unique linear fractional transformation.*

**Remark.** The assertion of uniqueness here reflects the fact a linear fractional transformation is, by definition, a mapping. As noted earlier, a representation in the form (2.1.4) is not unique.

## 2.2 The cross-ratio and mapping properties of linear fractional transformations

Consider the question of determining what complex-valued functions

$$F(z_1, z_2, \ldots, z_n), \qquad z_j \in \mathbb{S},$$

are invariant under all linear fractional transformations $f$:

$$F\left(f(z_1), f(z_2), \ldots f(z_n)\right) = F(z_1, z_2, \ldots, z_n), \quad \text{all } z_j \in \mathbb{S}.$$

It follows from Proposition 2.1.3 that if $n \leq 3$, such a function is constant:

$$F(z_1, z_2, \ldots, z_n) = F(0, 1, \infty).$$

The case $n = 4$ is more interesting. Given distinct $(z_1, z_2, z_3, z_4)$, we may choose the linear fractional transformation $g = g_{z_2,z_3,z_4}$ such that

$$g(z_2) = 0, \quad g(z_3) = 1, \quad g(z_4) = \infty. \tag{2.2.1}$$

Invariance implies

$$F(z_1, z_2, z_3, z_4) = F\left(g_{z_2,z_3,z_4}(z_1), 0, 1, \infty\right).$$

Thus the invariant functions $F : \mathbb{S}^4 \to \mathbb{C}$ are functions of $g_{z_2,z_3,z_4}(z_1)$. This latter function is termed the *cross-ratio*, denoted $[z_1, z_2, z_3, z_4]$. The prescription (2.2.1) gives

$$[z_1, z_2, z_3, z_4] = \frac{z_1 - z_2}{z_1 - z_4} \Big/ \frac{z_3 - z_2}{z_3 - z_4} = \frac{(z_1 - z_2)(z_3 - z_4)}{(z_1 - z_4)(z_3 - z_2)}. \tag{2.2.2}$$

We should check that the cross-ratio *is* invariant. Suppose that $z_1, z_2, z_3$, and $z_4$ are distinct points, and let $w_j = f(z_j)$. The cross-ratio

$$[w_1, w_2, w_3, w_4] = g(w_1),$$

where $g$ is the unique linear fractional transformation that maps the triple $(w_2, w_3, w_4)$ to $(0, 1, \infty)$. Then $g \circ f$ is the unique linear fractional transformation that maps the triple $(z_2, z_3, z_4)$ to $(0, 1, \infty)$, so

$$[z_1, z_2, z_3, z_4] = [g \circ f](z_1) = g(w_1) = [w_1, w_2, w_3, w_4].$$

The cross-ratio can be used to give a simple proof of an important geometric fact. By "line" here we mean straight line.

**Theorem 2.2.1.** *The image of a line or a circle under a linear transformation is either a line or a circle.*

Note that the assertion is not that lines go to lines and circles go to circles, but that each line goes to a line or a circle, and each circle goes to a line or a circle. Suppose for the moment that Theorem 2.2.1 is true, and consider the cross-ratio. If distinct points $z_1, z_2, z_3, z_4$ lie on a line or circle, then their images under the map $g$ of (2.2.1) should all lie on the real line, so the cross-ratio should be real. Conversely, if the cross-ratio is real and Theorem 2.2.1 is true, then the $z_j$ should all lie on a line or circle. This is, in fact, the case.

**Lemma 2.2.2.** *Suppose that $z_1, z_2, z_3$ are distinct points in the plane $\mathbb{C}$. The (unique) line or circle that they determine is the set of solutions $z$ of the equation*

$$\operatorname{Im}[z, z_1, z_2, z_3] = 0. \tag{2.2.3}$$

*Proof:* Let $g(z) = (az+b)/(cz+d)$ be the linear fractional transformation that takes $(z_1, z_2, z_3)$ to $(0, 1, \infty)$. Then (2.2.3) is equivalent to $\operatorname{Im} g(z) = 0$, i.e.

$$0 = (az+b)\overline{(cz+d)} - \overline{(az+b)}(cz+d) = i[\alpha|z|^2 + \bar{\beta}z + \beta\bar{z} + \gamma], \tag{2.2.4}$$

where

$$\alpha = 2\operatorname{Im}(a\bar{c}), \quad \beta = i(\bar{a}d - b\bar{c}), \quad \gamma = 2\operatorname{Im}(b\bar{d}).$$

If $\alpha = 0$ then (2.2.4) is the equation of a line. If $\alpha \neq 0$, then dividing by $\alpha$ and completing a square shows that (2.2.4) is equivalent to

$$\left|z + \frac{\beta}{\alpha}\right|^2 = \left|\frac{\beta}{\alpha}\right|^2 - \frac{\gamma}{\alpha},$$

the equation of a circle with center $-\beta/\alpha$. Since each $g(z_j)$ is one of $0, 1$, or $\infty$, the $z_j$ themselves lie on this line or circle. $\square$

The proof of Theorem 2.2.1 is now straightforward. Given a line or circle $\Gamma$ in the plane, choose three distinct points $z_j \in \Gamma$. Let $w_j$ be the image of $z_j$ under the linear fractional transformation $f$. By Lemma 2.2.2, $\Gamma$ is the set of solutions $z$ of

$$\operatorname{Im}[z, z_1, z_2, z_3] = 0. \tag{2.2.5}$$

By invariance of the cross-ratio, if $z$ satisfies (2.2.5), then $f(z)$ satisfies

$$\operatorname{Im}[f(z), w_1, w_2, w_3] = 0,$$

so $f(\Gamma)$ is the line or circle determined by $w_1, w_2, w_3$. □

A more geometric proof of Theorem 2.2.1 is outlined in Exercises 23 to 28.

An example of Theorem 2.2.1 is the map that takes the triple $(0,1,\infty)$ to $(-1,-i,1)$. This is the*Cayley transform*

$$c(z) = \frac{z-i}{z+i}. \tag{2.2.6}$$

The image of the real axis is the unit circle $\Gamma = \{z : |z| = 1\}$. Since $c(i) = 0$, the image of the upper half-plane $\mathbb{C}_+$ is the unit disk $\mathbb{D} = \{z : |z| < 1\}$. The image of the lower half-plane $\mathbb{C}_-$ is the complement of the closed unit disk. The inverse transform is

$$c^{-1}(w) = i\,\frac{1+w}{1-w}. \tag{2.2.7}$$

Under the Cayley transform, reflection across the real axis, i.e. the map $z \to \bar{z}$, corresponds to reflection across the unit circle $\Gamma$. In fact, setting $w = c(z)$, we have

$$w = c(z) \to c(\bar{z}) = \frac{1}{\overline{w}}. \tag{2.2.8}$$

## 2.3 Upper half plane and unit disk

There are two additional, widely used, geometric properties of linear fractional transformations.

**Proposition 2.3.1.** *The linear fractional transformation $f$ maps the upper half-plane $\mathbb{C}_+$ onto itself if and only if the coefficients in (2.1.4) can be chosen to be real, and $ad - bc > 0$.*

*Proof:* Suppose

$$f(z) = \frac{az+b}{cz+d}$$

has real coefficients. Then $f$ maps $\mathbb{R}$ to $\mathbb{R}$, so $f$ either preserves or interchanges the two components $\Omega_\pm$ of the complement. A calculation shows that the imaginary parts $\operatorname{Im} f(z)$ and $\operatorname{Im} z$ have the same sign if and only if $ad - bc > 0$; Exercise 19.

Conversely, suppose that $f$ above maps $\mathbb{C}_+$ onto itself. Since $f$ is bijective on $\mathbb{S}$, $f$ must map the complement to itself, and it follows that $f$ must map $\mathbb{R}\cup\{\infty\}$ to itself. If $c=0$, we may take $d=1$. Then since $f(0)$ and $f(1)$ are real, it follows that $a$ and $b$ are real.

If $c\neq 0$, we may take $c$ to be real. Since $f(\infty)$ is real, $a$ is real. Since the pole of $f$ must be real, $d$ must be real, and since $f(1)$ is real, $b$ must be real. As before, the condition $ad-bc>0$ is necessary. □

**Proposition 2.3.2.** *The linear fractional transformation $f$ maps the open unit disk $\mathbb{D}=\{z:|z|<1\}$ onto itself if and only if $f$ can be written in the form*

$$f(z) \;=\; \omega\,\frac{z-a}{\bar{a}z-1}, \qquad |a|<1,\;\; |\omega|=1. \tag{2.3.1}$$

*Proof:* Suppose $f$ has the form (2.3.1). Then if $|z|=1$,

$$|f(z)| \;=\; \left|\frac{z-a}{z(a\bar{z}-1)}\right| \;=\; \left|\frac{z-a}{a-z}\right| \;=\; 1.$$

It follows that $f$ maps the complement of the unit circle onto itself, and since $f(a)=0$ it follows that the bounded component $\mathbb{D}$ is mapped onto itself.

Conversely, suppose that $f$ maps $\mathbb{D}$ onto itself. Then $f$ must map the unit circle to itself. Thus if $|z|=1$, then

$$f(\bar{z})\overline{f(\bar{z})} \;=\; 1, \quad \text{so} \quad f(z^{-1}) \;=\; \frac{1}{\overline{f(\bar{z})}}. \tag{2.3.2}$$

Each side of the second equation defines a linear fractional transformation. Since these transformations agree on the unit circle, they must agree everywhere. Now $f$ has a zero at some point $a\in\mathbb{D}$, and (2.3.2) implies that $f$ has a pole at $1/\bar{a}$. Therefore $f$ has the form (2.3.1), with some constant $\omega$. The previous calculation shows that $\omega$ must have modulus 1. □

Proposition 2.3.1 and Proposition 2.3.2 can be strengthened, in part.

**Lemma 2.3.3.** **(Schwarz lemma)** *Suppose that $f:\mathbb{D}\to\mathbb{D}$ and $f(0)=0$. Then $|f(z)|\le|z|$, all $z\in\mathbb{D}$. In particular $|f'(0)|\le 1$ and equality holds if and only if $f$ is a rotation: $f(z)=\omega z$, $|\omega|=1$.*

*Proof:* Let $g(z)=f(z)/z$, $z\in\mathbb{D}$. Then $g$ is holomorphic and $|g(z)|\le 1/|z|$. The maximum modulus principle implies that $|g(z)|\le 1/r$ for $|z|\le r<1$. Taking the limit, $|g(z)|\le 1$, all $z\in\mathbb{D}$. If $|f'(0)|=1$, then the maximum value is taken at $z=0$, and the strong maximum modulus theorem says that $g$ is constant. □

**Theorem 2.3.4.** *Suppose that $f:\mathbb{D}\to\mathbb{D}$ or $f:\mathbb{C}_+\to\mathbb{C}_+$ is a holomorphic bijection. Then $f$ is a linear fractional transformation.*

*Proof:* Since the inverse Cayley transform (2.2.7) maps $\mathbb{D}$ onto $\mathbb{C}_+$, it is enough to consider the case $f : \mathbb{D} \to \mathbb{D}$. We may replace $f$ by its composition with a suitable linear fractional transformation, and assume that $f(0) = 0$. It is enough to show that $f$ is a rotation. Lemma 2.3.3 applies to both $f$ and to its inverse $h$, and implies that the derivatives of $f$ and its inverse at $z = 0$ must both have modulus 1. Therefore both are rotations. □

## Exercises

1. An *affine map* of $\mathbb{C}$ to itself has the form $f(z) = az + b$, $a \neq 0$. Given $z_1 \neq z_2$, there is a unique affine map $f$ such that $f(z_1) = 0$, $f(z_2) = 1$. Find a function $\langle z, z_1, z_2 \rangle$ with the property that a function of 3 variables that is invariant under affine maps is a function of $\langle z, z_1, z_2 \rangle$ (analogous to the cross-ratio $[z, z_1, z_2, z_3]$).
2. Prove that the stereographic projection $\pi : S \setminus \{N\} \to \mathbb{C}$ is *conformal*: if two smooth curves in $S$ cross at a point $P \neq N$, then the angle made by their tangents is the same as the angle made by the tangents to their images in $\mathbb{C}$ at the point $\pi(P)$. (Hint: if $(w(s), t(s))$ is a smooth curve in $S$, and $z(s) = \pi(w(s), t(s))$, then the derivative $\dot{z}$ has the form
$$\dot{z}(s) \;=\; f'(w(s), t(s)) \cdot w(s).)$$
3. Prove that the image under $\pi$ of a circle in $S$ is a line or a circle in $\mathbb{C}$. (Hint: a circle in $S$ is the intersection of $S$ and a plane in $\mathbb{C} \times \mathbb{R}$. The plane may be taken to be defined by an equation
$$\{(w,t) : \mathrm{Re}\,(\bar{a}w) + bt \;=\; c\}, \qquad |a|^2 + b^2 = 1,\;\; b \in \mathbb{R},\; 0 \leq c < 1.$$
This induces an equation for $\pi(w,t)$.)
4. Prove the distance relation (2.1.2). (A good starting point is the identity for points $(w_j, t) \in S$:
$$|w_1 - w_2|^2 + (t_1 - t_2)^2 \;=\; 2 - 2\mathrm{Re}\,(w_1 \overline{w}_2 + t_1 t_2).)$$
5. Determine the relation between the length of the chord that joins two points on the unit sphere $S$ and the length of the arc of the great circle between the two points.
6. Suppose the $\gamma : [0,1] \to S$ is a smooth curve. Based on the distance function (2.1.3), express the length of (the image of) $\gamma$ as an integral.
7. Use Exercise 6 to show that the length of the arc of a great circle determined by two points in $S$ is what it should be: the angle between the two radii. (It is sufficient to consider the length of an interval on the positive real axis.)
8. Based on the distance formula, the “element of area” in $\mathbb{C}$ at the point $z = x + iy$ should be
$$\frac{dx\,dy}{(1+|z|^2)^2}.$$
Show that the area of the sphere $S$ is $\pi$.

9. Suppose that $a$ and $b$ are complex numbers with $|a|^2 + |b|^2 = 1$. Prove that the map $f : \mathbb{S} \to \mathbb{S}$ given by
$$f(z) = \frac{az+b}{\overline{a} - \overline{b}z}$$
is an isometry: $d(f(z), f(w)) = d(z,w)$, for every pair $z, w \in \mathbb{S}$. (Thus $f$ corresponds to a rotation of the sphere.)
10. Show that if $f : \mathbb{S} \to \mathbb{S}$ is meromorphic, then $f$ is a rational function (i.e. a quotient of polynomials).
11. Verify that
$$\frac{az+b}{cz+d} = w$$
has a unique solution $z \in \mathbb{S}$ for each $w \in \mathbb{S}$ if and only if $ad - bc \neq 0$.
12. Prove Proposition 2.1.3.
13. Derive formula (2.2.2).
14. Verify (2.1.8).
15. Give another proof that $ad - bc \neq 0$ is necessary in Proposition 2.1.1 by computing the derivative of $f_A$.
16. Show that there is no loss of generality to assume that the matrix $A$ associated with a linear fractional transformation has determinant 1.
17. Show directly that the composition of two linear fractional transformations of the form (2.3.1) has the same form.
18. Verify that the transformations (2.2.6) and (2.2.7) map $\mathbb{C}_+$ to $\mathbb{D}$ and $\mathbb{D}$ to $\mathbb{C}_+$, respectively.
19. Show that if
$$f(z) = \frac{az+b}{cz+d}$$
has real coefficients, then the imaginary parts of $z$ and $f(z)$ have the same sign if and only if $ad - bc > 0$.
20. Show that a linear fractional transformation $f$ that is not the identity map has exactly one or two *fixed points*: points $z$ such that $f(z) = z$.
21. Suppose the linear fractional transformation $f$ has exactly one fixed point. Investigate the possible behaviors of the sequence $z_0 = z$, $z_n = f(z_{n-1})$ as $n \to \infty$. (Hint: the fixed point can be assumed to the point at $\infty$.)
22. Suppose the linear fractional transformation $f$ has exactly two fixed points. Investigate the possible behaviors of the sequence $z_0 = z$, $z_n = f(z_{n-1})$ as $n \to \infty$. (Hint: the fixed points can be assumed to be 0 and $\infty$.)
23. Show that each linear fractional transformation is either of the form (2.1.5) or has the form $f \circ r \circ g$, where $f$ and $g$ have the form (2.1.5) and $r(z) = 1/z$. (Hint: if the linear fractional transformation has a pole at a finite point, use $g$ to move the pole to the origin.)

Since linear fractional transformations of the form (2.1.5) map circles to circles and lines to lines, the preceding exercise shows that Theorem 2.2.1 can be proved more directly by showing that the inversion $r(z) = 1/z$ takes a circle or line to a circle or line. The next exercises construct such a proof.

24. Show that to prove Theorem 2.2.1, it is enough to prove that the inversion $r$ maps (a) lines not through the origin to circles, (b) circles not centered at the origin to circles.
25. Show that the two cases in the preceding exercise can be reduced to (a) the line $\{z : \operatorname{Im} z = 1\}$ and (b) a circle with center 1.
26. Suppose that $L$ is the line $\{z : \operatorname{Im} z = 1\}$. We can expect $r(L)$ to be symmetric with respect to reflection about the real axis. Since $r(\infty) = 0$ and $r(1) = 1$, if $r(L)$ is a circle it should have center $1/2$ and radius $1/2$. Show that for every real $t$, $r(1+it)$ lies on this circle.
27. Show that the preceding exercise implies that $r$ takes each circle that passes through the origin to a straight line. (Note that $r$ is its own inverse.)
28. Suppose that $C$ is a circle with center 1 and radius $\lambda \neq 1$. By symmetry, $r(C)$ should be symmetric with respect to reflection about the real axis. If the image is a circle, the images $r(1 \pm \lambda)$ should be the endpoints of a diameter of $r(C)$. Assuming this, compute the center and radius of the corresponding circle $C'$ and verify that each point of $r(C)$ lies on the circle $C'$.

## Remarks and further reading

Linear fractional transformations are a particularly important example of a group of geometric transformations; see Chapter 3 and the references there. For a general discussion of geometric transformations, see Kawakubo [74].

These transformations—especially various subgroups—also play a key role in the study of conformal mapping and modular forms; see Chapters 5, 6, and 17.

# Chapter 3
# Hyperbolic geometry




Euclidean plane geometry is based on primitive notions of "point" and "line," fleshed out with notions of "distance" and "congruence." Fundamental to the idea of congruence are the distance-preserving motions: translations, rotations, and combinations of these.

The euclidean axioms, or postulates, seemed self-evident, with the possible exception of the *parallel postulate*. One version of this postulate says: given a line $L$ and a point $P$ not on $L$, there is a unique line $L'$ through $P$ that does not intersect the line $L$. Much effort was expended in trying to derive this from the other, more obvious, postulates/axioms.

In the early 19th century, Lobachevsky produced a geometry that satisfied the other postulates of Euclid but violated the parallel postulate: given a point not on a line, there are many lines through that point that do not intersect the given line.

Poincaré modeled Lobachevsky's geometry in the upper half-plane $\mathbb{C}_+$. The equivalent construction in the unit disk $\mathbb{D}$ is the focus of this chapter.

## 3.1 Distance-preserving transformations and "lines"

To model Lobachevsky's geometry in $\mathbb{D}$, we start by declaring what the lines and the distance-preserving transformations will be. We then derive a distance function to fit.

The basic assumption is that the role played by translations and rotations in Euclidean plane geometry is to be played by the linear fractional transformations that map the unit disk $\mathbb{D}$ onto itself. We denote this group of transformations by $\mathrm{Aut}(\mathbb{D})$: the automorphism group of $\mathbb{D}$. Recall from Section 2.3 that these are the transformations $f$ that have the form

$$f(z) = \omega \frac{z-a}{\bar{a}z-1}, \qquad |a|<1, \quad |\omega|=1. \tag{3.1.1}$$



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_3

It follows that $f(0) = 0$ if and only if $f$ is a rotation: $f(z) = \omega z$.

We will use the following additional properties of these transformations; see Exercises 1 and 2.

**Proposition 3.1.1.** *(a) Given two distinct points $z_1$, $z_2$ in $\mathbb{D}$, there is a unique $f \in$ $\mathrm{Aut}(\mathbb{D})$ such that*

$$f(z_1) = 0, \qquad 0 < f(z_2) < 1. \tag{3.1.2}$$

*(b) Linear fractional transformations are conformal maps: if two smooth curves meet a point $z$ that is not a pole of the linear fractional transformation $f$, then the angle between their images at $f(z)$ is the same as the angle between the curves at $z$.*

Our next task is to determine the "lines." We start with an obvious choice of lines in $\mathbb{D}$: the diameters of $\mathbb{D}$. We assume, as in Euclid, that there is exactly one line through each pair of distinct points, and it follows that we have now accounted for *all* the lines that pass through the origin. Since we want invariance under $\mathrm{Aut}(\mathbb{D})$, the collection of lines through a given point $a \in \mathbb{D}$ must be the image of the diameters under any linear fractional transformation that takes the origin to $a$. Note that it is sufficient to consider images of the interval $(-1,1)$. Theorem 2.2.1 tells us that, for $a \neq 0$, each such image is either a diameter or a circular arc. A diameter meets the boundary of $\mathbb{D}$ in a right angle, so preservation of angles tells us that the image meets the boundary of $\mathbb{D}$ in a right angle. If follows that each such image is either itself a diagonal or is a circular arc that meets the boundary—the unit circle—at right angles. Note that each such circular arc is contained in the sector determined by the radii that meet the boundary at the same two points.

Conversely, suppose that $C$ is a circular arc that meets the unit circle in right angles. Choose $f \in \mathrm{Aut}(\mathbb{D})$ that maps two distinct points on $C$ to points in $(-1,1)$. Thus $f(C)$ is not contained in a proper sector, so $f(C)$ is the interval $(-1,1)$. Thus $C$ is the image of this diameter under $f^{-1}$.

Summing up, we have identified precisely the lines of our geometry.

**Proposition 3.1.2.** *The set of images of the diameter $(-1,1)$ of $\mathbb{D}$ is the set consisting of all diameters of $\mathbb{D}$, and all circular arcs in $\mathbb{D}$ that meet the boundary in right angles. This set is invariant under* $\mathrm{Aut}(\mathbb{D})$.

Note that with this definition, there are many lines through a given point that are parallel to a given line—see Figure 3.1.

## 3.2 Construction of a distance function

Next, we want to equip $\mathbb{D}$ with a *distance function* $\rho$. The most basic properties demanded of $\rho$ are

$$\rho(z,z) = 0, \qquad \rho(z_1,z_2) = \rho(z_2,z_1) > 0 \text{ if } z_1 \neq z_2. \tag{3.2.1}$$

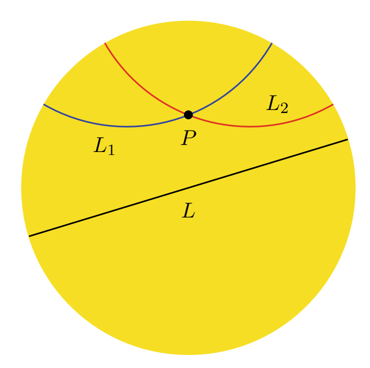



**Fig. 3.1** Two lines through $P$ parallel to $L$

We want the elements of $\mathrm{Aut}(\mathbb{D})$ to be distance-preserving:

$$\rho(f(z), f(w)) = \rho(z, w), \qquad \text{all } f \in \mathrm{Aut}(\mathbb{D}). \tag{3.2.2}$$

To pin down a scale, we require that $\rho$ look like the euclidean distance near $z = 0$:

$$\lim_{z \to 0} \frac{\rho(0, z)}{|z|} = 1. \tag{3.2.3}$$

We also want the lines for this geometry to be the *geodesics*: the shortest paths from one point to another. A characteristic of geodesics is *additivity of distance*: if distinct points $z_1$, $z_2$, and $z_3$ lie on the same line, and $z_2$ lies between $z_1$ and $z_3$, then we should have

$$\rho(z_1, z_3) = \rho(z_1, z_2) + \rho(z_2, z_3). \tag{3.2.4}$$

**Theorem 3.2.1.** *There is a unique function $\rho : \mathbb{D} \times \mathbb{D} \to \mathbb{R}$ that satisfies (3.2.1), has the invariance and scale properties (3.2.2) and (3.2.3), and has the additivity property (3.2.4) when the $z_j$ lie on a line and $z_2$ is between $z_1$ and $z_3$.*

*Proof:* We could establish the *existence* of such a function $\rho$ by writing down a formula and verifying (3.2.1)–(3.2.4). It is more illuminating to show how such a function can be constructed uniquely by first assuming these properties.

Suppose, then, that $\rho$ is such a function. Invariance under rotation implies that

$$\rho(0, z) = \rho(0, |z|). \tag{3.2.5}$$

We use invariance under linear fractional transformations to reduce (3.2.4) to a special case, and then determine the solution.

For convenience we abuse notation and write $\rho$ also as a function of $r$, $0 \le r < 1$:

$$\rho(0, z) = \rho(0, |z|) = \rho(|z|). \tag{3.2.6}$$

Given points $0 < r < r + \delta < 1$, we use the map

$$f_r(z) \;=\; \frac{z-r}{1-rz}. \tag{3.2.7}$$

Then

$$f_r(r) \;=\; 0, \qquad f_r(r+\delta) \;=\; \frac{\delta}{1-r(r+\delta)} \;=\; \frac{\delta}{1-r^2} + O(\delta^2) \tag{3.2.8}$$

as $\delta \to 0$. The additivity condition (3.2.4) becomes

$$\rho(r+\delta) \;=\; \rho(r) + \rho(r, r+\delta) \;=\; \rho(r) + \rho\left(\frac{\delta}{1-r(r+\delta)}\right). \tag{3.2.9}$$

We let $\delta \to 0+$ and use (3.2.8) and the normalization (3.2.3) to conclude that

$$\rho'(r) \;=\; \lim_{\delta\to 0} \frac{\rho(r+\delta)-\rho(r)}{\delta} \;=\; \frac{1}{1-r^2} \;=\; \frac{1}{2}\left[\frac{1}{1+r} + \frac{1}{1-r}\right]. \tag{3.2.10}$$

(Technically, we have only established this for a one-sided derivative, but the argument is easily adapted for $r-\delta$ in place of $r+\delta$.)

We want $\rho(0) = 0$, so we may integrate (3.2.10) to get

$$\rho(r) \;=\; \frac{1}{2}\log\left(\frac{1+r}{1-r}\right).$$

Using invariance, we may explicitly compute $\rho(z_1, z_2)$ for each pair $z_1$, $z_2$. The unique linear fractional transformation $f$ such that $f(z_1) = 0$ and $f(z_2) > 0$ has the form

$$f(z) \;=\; \omega \frac{z-z_1}{1-\bar{z}_1 z} \tag{3.2.11}$$

for suitable $\omega$ with $|\omega| = 1$. Some calculation yields

$$\rho(z_1, z_2) \;=\; \rho(|f(z_2)|) \;=\; \frac{1}{2}\log\frac{|1-\bar{z}_1 z_2| + |z_2 - z_1|}{|1-\bar{z}_1 z_2| - |z_2 - z_1|}, \tag{3.2.12}$$

Exercise 6.

In particular, for $z_1 = r$, $z_2 = s$, real, with $r < s$,

$$\rho(r,s) \;=\; \frac{1}{2}\log\frac{(1-rs)+|s-r|}{(1-rs)-|s-r|} \;=\; \frac{1}{2}\log\frac{(1-r)(1+s)}{(1+r)(1-s)}. \tag{3.2.13}$$

Clearly the symmetry and positivity conditions (3.2.1) are satisfied.

The argument to this point shows that there is at most one function $\rho$ that satisfies all the conditions. Let us show that $\rho$ given by (3.2.12) is invariant under $\mathrm{Aut}(\mathbb{D})$. Suppose that $z_1$, $z_2$ are distinct points of $\mathbb{D}$. We have defined

$$\rho(z_1, z_2) \;=\; \rho(f(z_2)),$$

where $f$ is the unique element of $\mathrm{Aut}(\mathbb{D})$ such that $f(z_1) = 0$ and $f(z_2) > 0$. Suppose that $g \in \mathrm{Aut}(\mathbb{D})$ and $g(z_j) = w_j$. Then $h = f \circ g^{-1}$ is the unique map in $\mathrm{Aut}(\mathbb{D})$ such

that $h(w_1) = f(z_1) = 0$ and $h(w_2) = f(z_2) > 0$. Therefore

$$\rho(g(z_1), g(z_2)) \;=\; \rho(h(w_1), h(w_2)) \;=\; \rho(f(z_1), f(z_2)) \;=\; \rho(z_1, z_2).$$

It remains to check the full additivity property. Using invariance, we may take the three points to be $-1 < r < s < t < 1$. Then (3.2.13) gives

$$\begin{aligned}\rho(r,s) + \rho(s,t) &= \frac{1}{2}\log\frac{(1+r)(1-s)(1+s)(1-t)}{(1-r)(1+s)(1-s)(1+t)} \\ &= \frac{1}{2}\log\frac{(1+r)(1-t)}{(1-r)(1+t)} \;=\; \rho(r,t). \qquad \square\end{aligned}$$

Note that $\rho(z_1, z_2) \to \infty$ if $|z_2| \to 1$: the lines in this geometry, as measured by $\rho$, are infinitely long.

## 3.3 The triangle inequality

The final steps in the analysis are to show that the distance function $\rho$ is a *metric* and that the lines are the geodesics. To show that $\rho$ is a metric, we need to verify the *triangle inequality*

$$\rho(z_1, z_3) \;\leq\; \rho(z_1, z_2) + \rho(z_2, z_3). \tag{3.3.1}$$

**Theorem 3.3.1.** *The distance function $\rho$ is a metric. In fact strict inequality holds in (3.3.1) unless the points $z_1, z_2, z_3$ lie in order on the same line.*

*Proof:* Note that the $\rho$-circles

$$C_r(z) \;=\; \{w \in \mathbb{D} :\; \rho(z, w) = r\}$$

are circles in the euclidean geometry. In fact the case $z = 0$ is (3.2.3). The general case follows from this because the linear fractional transformations in question preserve $\rho$ and map circles inside $\mathbb{D}$ to circles inside $\mathbb{D}$.

Consider (3.3.1). We may assume that $z_1$ and $z_3$ lie on the diameter $(-1, 1)$. If $\rho(z_1, z_2) > \rho(z_1, z_3)$, there is nothing to prove. If $\rho(z_1, z_2) = \rho(z_1, z_3)$, then $z_2$ lies on a circle all of whose points except $z_3$ itself have positive distance to $z_3$. Therefore the inequality is strict unless $z_2 = z_3$. Finally, suppose $\rho(z_1, z_2) = r < \rho(z_1, z_3) = t$. The additivity property implies that the $\rho$-circles

$$C_r(z_1), \quad C_{t-r}(z_3)$$

meet at a single point between $z_1$ and $z_3$ on the real line. By assumption $z_2$ lies on $C_r(z_1)$, so the inequality is strict unless $z_2$ lies on the real line, between $z_1$ and $z_3$. $\square$

The metric $\rho$ is called the *Lobachevsky metric* or the *hyperbolic metric* .

**Corollary 3.3.2.** *The lines in $\mathbb{D}$ are geodesics for the metric $\rho$, i.e. given two distinct points $z_1, z_2$ of $\mathbb{D}$, the shortest curve from $z_1$ to $z_2$ is the arc from $z_1$ to $z_2$ along the line that joins $z_1$ and $z_2$.*

*Proof:* Suppose that $\gamma : [a,b] \to \mathbb{D}$ is a continuous curve from $z_1$ to $z_2$. By definition, the length of $\gamma$ is the limit as $\max_j |x_{j+1} - x_j| \to 0$ of

$$\sum_{j=1}^{n} \rho(\gamma(x_{j-1}), \gamma(x_j)), \qquad a = x_0 < x_1 < \cdots < x_n = b.$$

In view of Theorem 3.3.1, this sum will be minimal if and only if each successive triple of points lies on a line. If so, then all the points lie on the same line and the sum is $\rho(z_1, z_2)$, independent of the partition of the interval $[a,b]$. □

## 3.4 Distance and area elements

The normalization (3.2.3) was chosen so that the Lobachevsky metric is asymptotically the euclidean metric near $z = 0$. To understand the behavior near another point $z \in \mathbb{D}$, we may use invariance under rotation and take $z = r > 0$. The metric is asymptotically a multiple of the euclidean metric in small neighborhoods of $r$. To compute the multiple we look at the distance from $r$ to $r + \varepsilon$ for small $\varepsilon > 0$:

$$\rho(r, r+\varepsilon) \;=\; \frac{\rho(r+\varepsilon) - \rho(r)}{\varepsilon} \cdot \varepsilon \;\approx\; \rho'(r) \cdot \varepsilon \;=\; \frac{\varepsilon}{1 - r^2}.$$

Thus, asymptotically near a point $z \in \mathbb{D}$, the Lobachevsky metric is the euclidean metric inflated by $1/(1 - |z|^2)$. In differential-geometric terms, it has a Riemannian metric

$$ds^2 \;=\; \frac{dx^2 + dy^2}{(1 - r^2)^2}. \tag{3.4.1}$$

It follows that the appropriate scaling of *area* is

$$\frac{dx\,dy}{(1 - r^2)^2} \;=\; \frac{r\,dr\,d\theta}{(1 - r^2)^2}. \tag{3.4.2}$$

In particular, the disk $\mathbb{D}$ itself has infinite area. However hyperbolic polygons in $\mathbb{D}$ have finite area—see Exercise 9.

Given that one can map $\mathbb{D}$ to the upper half-plane $\mathbb{C}_+$ and $\mathbb{C}_+$ to $\mathbb{D}$ by linear fractional transformations, it is clear that one can also model Lobachevsky geometry in $\mathbb{C}_+$. See the exercises starting with Exercise 10.

## Exercises

In the following exercises, "line" means an image of the real interval $(-1,1)$ under an element of $\mathrm{Aut}(\mathbb{D})$, i.e. a geodesic for $\rho$.

1. Prove that for distinct $z_1, z_2$ in $\mathbb{D}$ there is a unique $f \in \mathrm{Aut}(\mathbb{D})$ such that $f(z_1) = 0$ and $f(z_2) > 0$.
2. Show that every linear fractional transformation $f$ preserves angles: if two smooth curves $\gamma_1$, $\gamma_2$ cross at a point $z$ that is not a pole of $f$, then the angle between the tangent vectors to their images $f \circ \gamma_1$, $f \circ \gamma_2$ at $f(z)$ is the same as the angle between the tangent vectors to the curves at $z$. (Hint: $f'(z) \neq 0$.)
3. Given a line $L$ in $\mathbb{D}$ and a point $z \in \mathbb{D}$, not on $L$, show that there is a unique linear fractional transformation $f \in \mathrm{Aut}(\mathbb{D})$ such that $f(L)$ is the horizontal diameter $(-1,1)$ and $f(z)$ is on the positive imaginary axis.
4. Show that any two points in $\mathbb{D}$ are joined by a unique line. (In this and the exercises to follow, it helps to use invariance to reduce to a special case.)
5. If $L \subset \mathbb{D}$ is a line and $a \in \mathbb{D}$ is a point not on $L$, show that there are infinitely many lines that pass through $a$ and do not meet $L$. Thus this geometry is decidedly non-euclidean.
6. Verify the calculation leading to (3.2.12).
7. Show that, with the exception of $z = 0$, the euclidean center of the Lobachevsky circle $C_r(z)$ is not $z$: it lies closer to the boundary.
8. Show that the shortest distance from a point $z$ in $\mathbb{D}$ to a line $L$ is the distance from $z$ to $L$ along the (unique) geodesic through $z$ that meets $L$ in a right angle.
9. A (hyperbolic) *polygon* in $\mathbb{D}$ is a figure with vertices $z_1, z_2, \ldots z_n$, $n \geq 3$, that are distinct points of the closure of $\mathbb{D}$, and sides that are portions of lines in $\mathbb{D}$. Show that a polygon has finite hyperbolic area.
10. Use the Cayley transform and its inverse to construct a metric $\rho_+$ on $\mathbb{C}_+$ that is invariant under the linear fractional transformations that map $\mathbb{C}_+$ onto itself and has the positive imaginary axis as a geodesic. Here the standard normalization is
$$\lim_{z \to i} \frac{\rho_+(z,i)}{|z-i|} \;=\; 1. \tag{3.4.3}$$
11. What are the lines/geodesics in $\mathbb{C}_+$?
12. Give a direct construction of the metric $\rho_+$ of the preceding exercise by following the pattern used here for $\rho$. Use additivity and (3.4.3) to determine $\rho_+$ on the positive imaginary axis, and use invariance to define $\rho$ generally on $\mathbb{C}_+$.
13. Compute the local behavior of the Lobachevsky metric at a general point of $\mathbb{C}_+$.
14. What is the analogue for $\mathbb{C}_+$ of the Riemannian metric formula (3.4.1)?
15. What is the analogue for $\mathbb{C}_+$ of the area element (3.4.2)?
16. Show that a (hyperbolic) polygon in $\mathbb{C}_+$ has finite area.

## Remarks and further reading

Siegel [127] contains an efficient introduction to the subject. The half-plane model, in particular, plays a decisive role in geometry of three-manifolds; see Hubbard [68], Stahl [129], and Marden [97].

# Chapter 4
# Harmonic functions




Harmonic functions of two variables are closely related to holomorphic functions; in fact the real and imaginary parts of a holomorphic function are each harmonic. Conversely, at least locally, a real-valued harmonic function is the real part of a holomorphic function. A mean value property of these functions leads to analogues of the maximum modulus principle and its strong version.

Harmonic functions play a key role in the study of conformal mapping, via an important reflection property proved by Schwarz. The solution of the Dirichlet problem for a disk—finding a harmonic function with assigned values on the boundary of the disk—leads directly to important results for Fourier series and approximation, including two approximation theorems of Weierstrass.

Throughout this chapter we identify $\mathbb{R}^2$ with $\mathbb{C}$ in the usual way, and treat functions interchangeably in the form $u(x,y)$ or $u(x+iy)$.

## 4.1 Harmonic functions and holomorphic functions

A twice continuously differentiable function $u(x,y)$ defined on some open subset $\Omega$ of the plane is said to be *harmonic* if it satisfies *Laplace's equation*

$$u_{xx} + u_{yy} \;=\; 0. \tag{4.1.1}$$

Suppose $u$ and $v$ are real-valued, and set $f(x+iy) = u(x,y) + iv(x,y)$. Recall that the necessary and sufficient conditions for $f$ to be holomorphic are the Cauchy–Riemann equations

$$v_x \;=\; -u_y, \qquad v_y \;=\; u_x. \tag{4.1.2}$$

Suppose these equations are satisfied. Differentiating them shows that $u$ and $v$ are harmonic:

$$u_{xx} + u_{yy} \;=\; v_{yx} - v_{xy} \;=\; 0 \;=\; u_{xy} - u_{yx} \;=\; v_{xx} + v_{yy}.$$



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_4

**Proposition 4.1.1.** *The real part of a holomorphic function is harmonic. Conversely, if $u$ is a real-valued harmonic function in a disk $D = \{z : |z-a| < R\}$, then there is a function $f$ holomorphic in $D$ such that* $\operatorname{Re} f = u$.

*Proof:* The first statement has already been verified. For the converse, we want to construct a real-valued function $v$, defined in $D$, such that $u$ and $v$ satisfy the Cauchy–Riemann equations (4.1.2). After a translation, we may assume that $z_0 = 0$. We look for $v$ with $v(0) = 0$. Since $v(x,y)$ would be the value at $s = 1$ of $v(sx, sy)$, we should have

$$v(x,y) \;=\; \int_0^1 \frac{d}{ds}\{v(sx,sy)\}\,ds \;=\; \int_0^1 [xv_x(sx,sy) + yv_y(sx,sy)]\,ds.$$

If (4.1.2) is to hold, this would be the same as

$$v(x,y) \;=\; \int_0^1 [-xu_y(sx,sy) + yu_x(sx,sy)]\,ds. \tag{4.1.3}$$

Thus we define $v$ in $D$ by (4.1.3). Differentiating (4.1.3) with respect to $x$ and to $y$ shows that $v$ is a solution of (4.1.2); see Exercise 1. □

## 4.2 The mean value property, the maximum principle, and Poisson's formula

We need some properties of harmonic functions. One such property is the *mean value property*: the value at a point $z_0$ in the domain of definition of $u$ is the average value over nearby circles centered at that point:

$$u(z_0) \;=\; \frac{1}{2\pi}\int_0^{2\pi} u(z_0 + re^{i\theta})\,d\theta. \tag{4.2.1}$$

In fact, by Proposition 4.1.1, $u$ is the real part of a function $f$ that is holomorphic in a disk centered at $z_0$. Taking the real part of the Cauchy integral formula gives the result—see (1.2.4).

**Corollary 4.2.1.** (Maximum principle) *If a real function is harmonic on a bounded domain $\Omega$ and continuous on the closure of $\Omega$, then its minimum and maximum values are attained on the boundary.*

*Proof:* The value of $u$ at a point $z_0 \in \Omega$ lies between the minimum and maximum values of $u$ on nearby circles. Follow shrinking overlapping circles along a curve leading to the boundary to find a sequence of values $\le u(z_0)$ and a sequence of values $\ge u(z_0)$. □

**Corollary 4.2.2.** (Strong maximum principle) *Suppose that $u$ is a real function, harmonic in a connected open domain $\Omega$. If $u$ attains a maximum or minimum at a point of $\Omega$, then $u$ is constant.*

*Proof:* Suppose that an extreme value occurs at a point $z_0 \in \Omega$. It follows from (4.2.1) that $u$ has constant value $u(z_0)$ on every small disk centered at $z_0$; see Exercise 8. Since $\Omega$ is assumed to be connected, the union of all disks on which $u$ has this same value is all of $\Omega$. □

Formula (4.2.1) gives a very partial analogue, for harmonic functions, of Cauchy's formula. Let us look for a fuller analogue: a formula that expresses the value at each point of a disk as an integral over the circle that bounds the disk. In fact we want to do more—to solve the *Dirichlet problem* for the disk: given a continuous function $f$ on the boundary, find a function $u$, harmonic on the disk, that has $f$ as limit on the boundary. We assume for convenience that the disk is the unit disk $\mathbb{D}$. The basic idea is to expand the (desired) function in terms of a nice family of harmonic functions: the monomials $z^n$ and their complex conjugates $\bar{z}^n$:

$$\begin{aligned} u(z) &= a_0 + \sum_{n=1}^{\infty} a_n z^n + \sum_{n=1}^{\infty} a_{-n} \bar{z}^n \\ &= \sum_{n=-\infty}^{\infty} a_n r^{|n|} e^{in\theta}, \qquad |z| = r < 1. \end{aligned}$$

In particular, on the boundary $r = 1$ we want, at least formally,

$$f(\theta) = u(e^{i\theta}) = \sum_{n=-\infty}^{\infty} a_n e^{in\theta}. \tag{4.2.2}$$

The question is: how to choose the $\{a_n\}$? It is easily verified that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{in\theta} e^{-im\theta} \, d\theta = \begin{cases} 1 & \text{if } n = m; \\ 0 & \text{if } n \neq m. \end{cases} \tag{4.2.3}$$

Therefore a formal calculation, integrating (4.2.2) against $e^{-in\theta}$, gives

$$a_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\theta) e^{-in\theta} \, d\theta.$$

Let us define $a_n$ by this integral. Then $|a_n| \leq \sup |f(\theta)|$, so the series

$$\sum_{-\infty}^{\infty} a_n r^{|n|} e^{in\theta}$$

converges uniformly for $0 \leq r \leq 1 - \delta$, $\delta > 0$. The derivatives of any order also converge uniformly on these smaller disks, confirming that $u$ is harmonic. Inserting the definition of $a_n$, we have

$$u(re^{i\theta}) \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\varphi)\,P_r(\theta-\varphi)\,d\varphi, \tag{4.2.4}$$

where the *Poisson kernel* $P_r(\theta)$ is defined for $0 \le r < 1$ by

$$\begin{aligned} P_r(\theta) &= \sum_{n=-\infty}^{\infty} r^{|n|}(e^{i\theta})^n \\ &= \sum_{n=0}^{\infty}(re^{i\theta})^n + \sum_{n=0}^{\infty}(re^{-i\theta})^n - 1 \\ &= \frac{1}{1-re^{i\theta}} + \frac{1}{1-re^{-i\theta}} - 1 \\ &= \frac{1-r^2}{1-2r\cos\theta+r^2} \;>\; 0. \end{aligned} \tag{4.2.5}$$

The first formula for $P_r$, together with (4.2.3), implies that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} P_r(\theta)\,d\theta \;=\; 1. \tag{4.2.6}$$

Moreover, given $0 < \delta < 1$, if $\cos\theta \le 1-\delta$ then

$$P_r(\theta) \;\le\; \frac{1-r^2}{1-2r(1-\delta)+r^2} \;=\; \frac{1-r^2}{(1-r)^2+2r\delta} \;\le\; \frac{1-r^2}{2r\delta}. \tag{4.2.7}$$

Thus, as $r \to 1$, $P_r(\theta)$ is more and more highly concentrated around $\theta = 0$. Because of this and uniform continuity of $f$, it follows that $u_r(\theta) = u(re^{i\theta})$ converges uniformly to $f(\theta)$ as $r \to 1$; see Exercise 10.

Summarizing, we have the following.

**Theorem 4.2.3.** *Suppose that $f$ is a continuous function on the unit circle. Then there is a unique function $u$, harmonic in the unit disk $\mathbb{D}$, continuous on the closure, and equal to $f$ on the circle. For $z = re^{i\theta} \in \mathbb{D}$, $u$ is given by the formula*

$$u(re^{i\theta}) \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\varphi)\,P_r(\theta-\varphi)\,d\varphi, \tag{4.2.8}$$

*where $P_r$ is given by (4.2.5).*

(Uniqueness is a consequence of the maximum principle.)

Translating and rescaling, an analogous result holds for any disk: Exercise 5.

## 4.3 The Schwarz reflection principle

A consequence of Theorem 4.2.3 that is very important for the theory of conformal mapping is the *Schwarz reflection principle.* It comes in two forms—and in one more form in the next chapter.

**Theorem 4.3.1.** *Suppose that $\Omega$ is a domain that is invariant under $z \to \bar{z}$, and that $I = \Omega \cap \mathbb{R}$ is not empty. Suppose that u is harmonic in $\Omega_+ = \Omega \cap \mathbb{C}_+$, the intersection of $\Omega$ with $\{z : \operatorname{Im} z > 0\}$ and that*

$$\lim_{z \in \Omega_+, z \to I} u(z) = 0.$$

*Then the extension of u to $\Omega$ that is defined on I and on $\Omega_- = \Omega \cap \{z : \operatorname{Im} z < 0\}$ by*

$$u(z) = 0, \quad z \in I, \qquad u(z) = -u(\bar{z}), \quad z \in \Omega_-, \tag{4.3.1}$$

*is harmonic in $\Omega$.*

*Proof:* Calculation of the derivatives shows that the extension is harmonic on $\Omega_-$, so we need only show that it is harmonic near points of $I$. Suppose $z_0 \in I$ and suppose that the closure of the disk $D = D_\rho(z_0) = \{z : |z - z_0| < \rho\}$ of radius $\rho$ is contained in $\Omega$. There is a function $u^*$, harmonic in $D$, that agrees with $u$ on the boundary of $D$. It can be checked from the formula (4.2.8), adapted to $D$, that the condition (4.3.1) on the circle that bounds $D$ implies that $u^*$ satisfies (4.3.1) throughout $D$. In particular, $u^* = 0$ on the intersection of $D$ with $I$. Thus $u^*$ agrees with $u$ on the boundary of the upper half-disk $D_+$, so $u^*$ agrees with $u$ on $D_+$, and thus on the entire disk $D$. This shows that $u$ is harmonic near each point of $I$. □

Theorem 4.3.1 allows for a strengthening of the reflection principle for holomorphic functions, Theorem 1.6.1:

**Theorem 4.3.2.** *Let $\Omega$ be as in Theorem 4.3.1. Suppose that $f$ is holomorphic in $\Omega_+$, and that*

$$\operatorname{Im} f(z) \to 0 \quad \text{as } z \to \mathbb{R}, \; z \in \mathbb{C}_+.$$

*Then $f$ extends to be holomorphic on all of $\Omega$ by taking*

$$f(z) = \overline{f(\bar{z})}, \qquad z \in \Omega_-. \tag{4.3.2}$$

*Proof:* It is enough to prove the result in the case of a disk $D$ with center on the real axis. Let $u$ be the imaginary part of $f$. By Theorem 4.3.1, $u$ extends across $\mathbb{R}$ to the lower half of $D$. By Proposition 4.1.1, $u$ is the real part of a function $g$ that is holomorphic in $D$. In the upper half-disk, $u$ is the real part of $-if$. It follows that $f - ig$ is a real constant $c$ on the upper half disk. Therefore $ig + c$ is the extension of $f$ to the disk. Since $ig + c$ is real on $D \cap \mathbb{R}$, it follows from the standard reflection principle, Theorem 1.6.1, that (4.3.2) is the extension of $f$. □

## 4.4 Application: approximation theorems

Another consequence of the argument that led to Theorem 4.2.3 is a simple proof of one of two approximation theorems of Weierstrass; see Exercise 11.

**Theorem 4.4.1.** *(Weierstrass) Suppose that $f : \mathbb{R} \to \mathbb{C}$ is a continuous complex-valued function with period $2\pi$: $f(x+2\pi) = f(x)$. Then $f$ can be approximated uniformly by trigonometric polynomials, i.e. functions of the form*

$$g(x) \;=\; \sum_{k=-n}^{n} a_k e^{ikx}.$$

The other well-known Weierstrass approximation theorem can be deduced from Theorem 4.4.1; see Exercise 12.

**Theorem 4.4.2.** *(Weierstrass) Suppose that $f$ is a continuous complex-valued function defined on a bounded real interval $[a,b]$. Then $f$ can be approximated uniformly by polynomials.*

Theorem 4.4.1 also allows us to flesh out some of the discussion of Hilbert spaces in Chapter 1. In fact it leads to a natural relationship between the two spaces introduced there: the space $l^2(Z)$ of two-sided complex sequences $\mathbf{x} = \{x_n\}_{n=-\infty}^{\infty}$ with inner product

$$(\mathbf{x},\mathbf{y}) \;=\; \sum_{n=-\infty}^{\infty} x_n \bar{y}_n,$$

and the space $L^2_{\text{per}}(\mathbb{R})$. The latter space is the completion of the space of continuous functions $u : \mathbb{R} \to \mathbb{C}$ that are periodic with period $2\pi$ with respect to the metric induced by the inner product

$$(u,w) \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi} u(x)\overline{w(x)}\,dx.$$

The orthogonality property (4.2.3) says that the functions

$$\varphi_n(x) \;=\; e^{inx}, \qquad n = 0, \pm 1, \pm 2, \ldots \tag{4.4.1}$$

are an orthonormal set in $L^2_{\text{per}}(\mathbb{R})$.

**Theorem 4.4.3.** *(Riesz–Fischer) The functions (4.4.1) are an orthonormal basis for the space $L^2_{\text{per}}(\mathbb{R})$. Therefore the mapping*

$$u \;\to\; \{\hat{u}(n)\}_{n=-\infty}^{\infty}, \qquad \hat{u}(n) \;=\; (u, \varphi_n), \tag{4.4.2}$$

*is a bijective map from $L^2_{\text{per}}(\mathbb{R})$ to $l^2(\mathbb{Z})$ that preserves the inner product.*

*Proof:* By definition, if $u$ belongs to the completion $L^2_{\text{per}}(\mathbb{R})$, then given $\varepsilon > 0$, there is a continuous periodic function $v$ such that $||u-v|| < \varepsilon/2$. In view of Theorem 4.4.1, there is an integer $n$ and an element $w$ in $H_n$, the span of $\{\varphi_j, |j| \le n\}$, such that $||v-w|| < \varepsilon/2$. Therefore $||u-w|| < \varepsilon$. By Bessel's inequality (1.9.5), the element in $H_n$ closest to $u$ is

$$u_n = \sum_{j=-n}^{n} (u, \varphi_j)\varphi_j = \sum_{j=-n}^{n} \hat{u}(j)\,\varphi_j.$$

Therefore $||u-u_n|| < \varepsilon$. We have shown that the sequence $\{u_n\}$ converges to $u$, so $\{\varphi_n\}$ is a basis. Bessel's equality (1.9.6) gives

$$||u||^2 = \sum_{n=-\infty}^{\infty} |\hat{u}(n)|^2.$$

Thus the sequence $\{\hat{u}(n)\}_{-\infty}^{\infty}$ belongs to $l^2(\mathbb{Z})$ and has the same norm as $u$. We have shown that the map (4.4.2) is injective.

Conversely, if the sequence $a = \{a_k\}_{-\infty}^{\infty}$ belongs to $l^2(\mathbb{Z})$, set

$$u_n = \sum_{k=-n}^{n} a_k\,\varphi_k. \tag{4.4.3}$$

Then $\{u_n\}$ is a Cauchy sequence in $L^2_{\text{per}}(\mathbb{R})$. It follows from the Cauchy–Schwarz inequality (1.9.4) that the *Fourier coefficients* $\hat{u}(k)$ of the limit $u$ are

$$\hat{u}(k) = \lim_{n\to\infty} \hat{u}_n(k). \tag{4.4.4}$$

But $\hat{u}_n(k) = a_k$ for each $n \ge |k|$. Therefore the map (4.4.2) is surjective. It is easily checked that

$$(u,w) = \sum_{n=-\infty}^{\infty} \hat{u}(n)\,\overline{\hat{w}(n)}.$$

□

## Exercises

1. Verify that if $v$ is defined by (4.1.3), then the Cauchy–Riemann equations are satisfied.
2. Suppose that $\Omega$ is simply connected, and suppose that $u : \Omega \to \mathbb{R}$ is harmonic. Show that there is $f : \Omega \to \mathbb{C}$ such that $f$ is holomorphic and $\operatorname{Re} f = u$. Is $f$ unique?
3. Suppose that $f : \Omega_1 \to \Omega_2$ is holomorphic and $u : \Omega_2 \to \mathbb{R}$ is harmonic. Prove that the composition $u \circ f$ is harmonic.

4. Suppose that $u$ is real-valued, harmonic on $\mathbb{C}_+$, and continuous on the closure $\mathbb{C}_+ \cup \mathbb{R}$. Suppose also that $u \geq 0$ on $\mathbb{R}$. (a) Show that $u$ is not necessarily non-negative on $\mathbb{C}_+$.

   (b) Show that if $u(z)$ has a limit as $z \to \infty$, then $u$ is non-negative on $\mathbb{C}_+$.
5. Suppose that $f$ is a continuous function on the circle $\{z : |z - z_0| = R\}$. Find an explicit formula for a function $u$ that is harmonic in $\{z : |z - z_0| < R\}$ and converges to $f$ on the boundary.
6. (Harnack's inequality) Suppose that $f$ in Exercise 5 is real-valued.

   (a) Prove that for $|z - z_0| = r < R$,

$$|u(z)| \;\leq\; \frac{1}{2\pi}\frac{R+r}{R-r}\int_0^{2\pi} |u(Re^{i\theta})|\,d\theta.$$

   (b) Suppose that $f \geq 0$. Prove that for $|z - z_0| = r < R$,

$$\frac{1}{2\pi}\frac{R-r}{R+r}\int_0^{2\pi} u(Re^{i\theta})\,d\theta \;\leq\; u(z).$$

   (c) Deduce from (a) and (b) that $f \geq 0$ implies that if $|z - z_0| = r < R$ then

$$\frac{R-r}{R+r}u(z_0) \;\leq u(z) \;\leq\; \frac{R+r}{R-r}u(z_0).$$

7. (Harnack's principle) Suppose that $\{u_n\}$ is a non-decreasing sequence of functions harmonic in a domain $\Omega$. Prove that either $u_n$ converges to a harmonic function $u$, uniformly on each compact subset of $\Omega$, or $u_n$ converges to $\infty$, uniformly on each compact subset of $\Omega$.
8. Prove the point that is taken for granted in the proof of Corollary 4.2.2: suppose $u$ is continuous on a circle $|z - z_0| = r$ and has values $\leq a$ on the circle. If the mean value

$$\frac{1}{2\pi}\int_0^{2\pi} u(z_0 + re^{i\theta})\,d\theta \;=\; a, \tag{4.4.5}$$

   then $u \equiv a$ on the circle.
9. Check that if $f$ in Theorem 4.2.3 satisfies $f(-\theta) = -f(\theta)$, then $u(\bar{z}) = -u(z)$.
10. Use (4.2.6) and (4.2.7) to fill in the details in the proof that the function $u$ defined by (4.2.8) is the solution to the Dirichlet problem with boundary value $f$, i.e. that $u_r$ converges uniformly to $f$ as $r \to 1$.
11. Use the preceding exercise and the uniform convergence of the series for $u_r$ to prove Theorem 4.4.1.
12. Use Theorem 4.4.1 to prove Theorem 4.4.2. (Note that the interval may be rescaled and the function $f$ extended, and also that convergence is only asked for over some bounded interval.)
13. Fill in some details in the proof of Theorem 4.4.3: show that the sequence (4.4.3) is a Cauchy sequence, and that the $a_k$ are the coefficients for the limit $u$.

14. A periodic function $f$ is said to have an *absolutely convergent* Fourier series if

$$||f||_a \equiv \sum_{n=-\infty}^{\infty} |a_n| < \infty, \qquad a_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\, e^{-inx}\, dx.$$

(a) Suppose this is the case. Prove that the series

$$\sum_{n=-\infty}^{\infty} a_n e^{inx} \tag{4.4.6}$$

converges uniformly to $f$. Deduce that $f$ is bounded and uniformly continuous.

(b) Suppose also that $f'$ is bounded. Prove that $|a_0| \leq \sup|f(x)|$ and $\sum_{n\neq 0} |a_n| \leq \sup|f'(x)|$.

15. A famous theorem of Wiener says that if $f$ has an absolutely convergent Fourier series and if, for each $x$, $f(x) \neq 0$, then $1/f$ has an absolutely convergent Fourier series. This exercise relies on Exercise 14 and sketches a proof due to Newman [106].

(a) We may suppose that $f(x) \geq 1$, all $x$. Let $P$ be a partial sum of the series (4.4.6) chosen so that $||P - f||_a < 1/3$. Let

$$S = \sum_{n=0}^{\infty} \frac{(P-f)^n}{P^{n+1}}.$$

Prove that the series converges in the norm $||\ ||_a$, and that the limit has an absolutely convergent Fourier series.

(b) Prove that $S = 1/f$.

## Remarks and further reading

This chapter merely touches on some large areas of real analysis. For more on harmonic functions in $\mathbb{R}^n$, see Axler, Bourdon, and Ramey [14]. Maximum principles for solutions of partial differential equations are treated by Protter and Weinberger [117] and Pucci and Serrin [118]. The classic treatise for Fourier series is Zygmund [149]. There are many textbooks on Fourier analysis, e.g. Grafakos [52].

In complex analysis, harmonic functions are particular cases of subharmonic functions (in one or several complex variables) and plurisubharmonic functions (several complex variables); see Hayman and Kennedy [63], Hayman [62], and texts on several complex variables, such as Krantz [79], Ohsawa [110], and Hörmander [66].

The Poisson kernel is one example of the important concept of an *approximate identity*; see Section 18.1.

# Chapter 5
# Conformal maps and the Riemann mapping theorem




A conformal map is one that preserves angles. In the case of mappings from one connected domain in $\mathbb{C}$ to another, such a map is holomorphic, or else its complex conjugate is holomorphic. In this chapter we consider the question of existence of an invertible holomorphic map from one domain to another. The fundamental result is Riemann's theorem: a simply connected plane domain that is not the whole plane can be mapped bijectively onto the unit disk (or, equivalently, onto $\mathbb{C}_+$). The proof relies on some clever uses of linear fractional transformations, together with an important compactness result (for functions) of Ascoli–Arzelá.

The proof of the general form of the Riemann mapping theorem is not constructive. When the domain $\Omega$ is a polygon, the Schwarz–Christoffel theorem establishes the form of the inverse map from $\mathbb{C}_+$ to $\Omega$.

In the final section of the chapter we touch on the subject of the inverses of Riemann maps: conformal maps from the unit disk to a domain in $\mathbb{C}$. The first steps on the long road to de Brange's proof of the Bieberbach conjecture are presented.

## 5.1 Conformal maps

A map $F$ from an open subset $\Omega$ of $\mathbb{R}^2$ into $\mathbb{R}^2$ is said to be *conformal* if, for each pair of smooth curves that pass through a point $P$ of $\Omega$, the angle between the tangents of the image curves at $F(P)$ is the same as the angle between the tangents of the original curves at $P$. Write $F(x,y) = (u(x,y), v(x,y))$; the functions $u$ and $v$ are assumed to have continuous first partial derivatives. It follows from the chain rule that the map from tangent vector to tangent vector, written as column vectors, is given by the matrix

$$A = \begin{bmatrix} u_x & v_x \\ u_y & v_y \end{bmatrix}. \tag{5.1.1}$$

In order for the images of non-zero angles to be well-defined, it is necessary and sufficient that $A$ be invertible. Let us assume for now that $A$ preserves the direction



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_5

of angles, as well as their magnitude. Then $\det A > 0$. Suppose first that the basis vector $(1,0)^t$ is mapped to a multiple of itself, $(\lambda_1,0)^t$. Then $(0,1)^t$ is mapped to some $(0,\lambda_2)^t$ and $(1,1)^t$ is mapped to $(\lambda_1,\lambda_2)^t$. Preservation of angles and directions implies that $\lambda_2 = \lambda_1$. It follows that $\lambda_1^2 = \det A$ and $A^tA = (\det A)I$, where $I$ is the identity matrix.

In the general case we may compose $A$ with a rotation matrix

$$B = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}$$

so that $BA$ maps $(1,0)^t$ to a multiple of itself. Since $BA$ preserves angle and direction, the preceding argument shows that $(BA)^tBA = \det(BA)\,I$. Note that $B^tB = I$, so

$$(\det A)I = \det(BA)I = (BA)^tBA = A^tB^tBA = A^tA.$$

Thus $A^t = (\det A)A^{-1}$:

$$\begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix} = \begin{bmatrix} v_y & -v_x \\ -u_y & u_x \end{bmatrix}.$$

and we obtain the Cauchy–Riemann equations $u_x = v_y$, $u_y = -v_x$. Thus

$$f(x+iy) = u(x,y) + iv(x,y)$$

is holomorphic for $(x,y) \in \Omega$. The same argument when $A$ reverses the direction of angles leads to the equations

$$u_x = -v_y, \quad u_y = v_x.$$

These equations imply that

$$\overline{f(x+iy)} = u(x,y) - iv(x,y)$$

is holomorphic. Thus, identifying $\mathbb{R}^2$ with the complex plane, we find that a conformal map is either a holomorphic map with non-vanishing derivative, or such a holomorphic map followed by complex conjugation. It is enough to study the holomorphic case.

## 5.2 The Riemann mapping theorem

Given two domains in the complex plane, a natural question is whether there exists a bijective holomorphic map from one onto the other. Linear fractional transformations are conformal maps: Exercise 2 of Chapter 3. Therefore we know that a half plane can be mapped conformally onto a disk, or onto the region outside a circle in the Riemann sphere $\mathbb{S}$, and conversely. The function $\sqrt{z}$ maps the right half plane to a wedge with angle $\pi/2$, and so on. On the other hand, a holomorphic map from the entire complex plane to a disk must be constant (Liouville's theorem). Moreover it

is easily seen that a bijective holomorphic image of a simply connected domain is simply connected. Therefore neither the entire plane nor the plane minus a single point can be mapped to the unit disk.

The definitive result is the following.

**Theorem 5.2.1. (Riemann mapping theorem)** *If $\Omega$ is an open, simply connected, proper subset of the plane, then there is a bijective holomorphic map $f$ that maps $\Omega$ onto the unit disk $\mathbb{D}$.*

*Given a point $z_0 \in \Omega$, we may specify $f(z_0) = 0$, $f'(z_0) > 0$, and these conditions determine $f$ uniquely.*

The proof involves a number of steps. The first step is a reduction to the case of bounded $\Omega$. Choose $a$ not in $\Omega$. Then $z - a$ is never zero on the simply connected domain $\Omega$, so we may choose a branch of $\sqrt{z-a}$ that is holomorphic on $\Omega$, see Section 1.4. This branch maps $\Omega$ bijectively onto a domain $\Omega_1$. Choose a point $b \in \Omega_1$. For some $\varepsilon > 0$, the disk $\{z : |z-b| < \varepsilon\}$ is contained in $\Omega_1$. If $z$ is in $\Omega_1$ then $-z$ is not, so $\Omega_1$ lies outside the disk $\{z : |z+b| < \varepsilon\}$. The map $z \to 1/(z+b)$ takes $\Omega_1$ bijectively onto a bounded domain $\Omega_2$. Thus we may replace $\Omega_1$ by $\Omega_2$, and assume that $\Omega$ itself is bounded.

For the next step, choose a point $z_0 \in \Omega$ and let $\mathscr{F}$ be the family of bijective holomorphic maps $f$ from $\Omega$ into the unit disk $\mathbb{D}$ such that $f(z_0) = 0$ and $f'(z_0) > 0$. Note that this family is not empty: $f(z) = \varepsilon(z - z_0)$ will belong to $\mathscr{F}$ if $\varepsilon > 0$ is small enough.

**Lemma 5.2.2.** *The family $\mathscr{F}$ contains a function $f$ such that $f'(z_0) \geq g'(z_0)$ for each $g \in \mathscr{F}$.*

Assuming this lemma, let us prove the Riemann mapping theorem. Suppose the function $f$ of Lemma 5.2.2 omits a point $a \in \mathbb{D}$. Suppose first, for simplicity, that $a > 0$. A branch of the square root can be chosen so that

$$g(z) = \sqrt{\frac{z-a}{az-1}}$$

is holomorphic on $f(\Omega) \subset \mathbb{D}$. The linear fractional transformation under the radical sign maps $\mathbb{D}$ to $\mathbb{D}$, so the composition $g \circ f$ maps $\Omega$ into $\mathbb{D}$. Note that $g(0) = \sqrt{a}$. Let

$$h(z) = \frac{z - \sqrt{a}}{\sqrt{a}z - 1}, \tag{5.2.1}$$

and let $f_1 = h \circ g \circ f$. Then $f_1$ is bijective from $\Omega$ into $\mathbb{D}$, and

$$f_1'(z_0) = h'(\sqrt{a})\, g'(0)\, f'(z_0). \tag{5.2.2}$$

But

$$g'(z) = \frac{1}{2g(z)} \frac{(az-1)-a(z-a)}{(az-1)^2} = \frac{1}{2g(z)} \frac{a^2-1}{(az-1)^2},$$

and

$$h'(z) = \frac{(\sqrt{a}z-1)-\sqrt{a}(z-\sqrt{a})}{(\sqrt{a}z-1)^2} = \frac{a-1}{(\sqrt{a}z-1)^2},$$

so

$$g'(0) = \frac{a^2-1}{2\sqrt{a}}, \qquad h'(\sqrt{a}) = \frac{1}{a-1}$$

and

$$f_1'(z_0) = \frac{a+1}{2\sqrt{a}} f'(z_0).$$

But since $0 < a < 1$ we have $a+1-2\sqrt{a} = (1-\sqrt{a})^2 > 0$, so $f_1'(z_0) > f'(z_0)$, contradicting the assumption that $f'(z_0)$ is maximal.

The preceding argument assumed that $f(\Omega)$ omitted a point $a \in \mathbb{D}$ and that $a > 0$. Otherwise, we may assume that the omitted point has the form $\omega a$, where $|\omega| = 1$ and $a > 0$, and take

$$f_1(z) = \omega h(g(\bar{\omega} f(z)))$$

with $g$ and $h$ defined as before. Again we find that $f_1'(z_0) > f'(z_0)$, a contradiction. This contradiction shows that each function $f$ of Lemma 5.2.2 maps onto $\mathbb{D}$.

To complete the proof that Lemma 5.2.2 implies the Riemann mapping theorem, we need to prove uniqueness. The key result here is a special case of Theorem 2.3.4.

**Lemma 5.2.3.** *Suppose that $f : \mathbb{D} \to \mathbb{D}$ is a holomorphic bijection with $f(0) = 0$. Then $f$ is a rotation.*

Suppose now that $f$ and $g$ are two mappings that have the properties of the map in Lemma 5.2.2. Then $h = g \circ f^{-1}$ maps $\mathbb{D}$ onto $\mathbb{D}$ and $h(0) = 0$. By Schwarz's lemma, $h$ is a rotation. But $h$ has derivative $= 1$ at $z = 0$, so $h$ is the identity. Thus $g = f$. This completes the argument that Lemma 5.2.2 implies Theorem 5.2.1.

## 5.3 Proof of Lemma 5.2.2; the Ascoli–Arzelà theorem

The first step is to establish some properties of the family $\mathscr{F}$. Define $\Omega_m$, $m = 1, 2, \ldots$ to be the subset of $\Omega$ consisting of points $z$ whose distance to the boundary is $> 1/m$, i.e. the closed disk $\{w : |w-z| \le 1/m\}$ is contained in $\Omega$. The closure $\overline{\Omega_m}$ is compact.

For $z$ in $\Omega_m$ and $f$ in $\mathscr{F}$, we have the estimate

$$|f'(z)| = \left| \frac{1}{2\pi i} \int_{|\zeta - z| = 1/2m} \frac{f(\zeta)}{(\zeta - z)^2} \, d\zeta \right| \le 2m; \qquad (5.3.1)$$

see Exercise 1. Thus, restricted to $\overline{\Omega}_m$, the functions $\mathscr{F}$ are *uniformly bounded* (in fact $|f(z)| < 1$) and *uniformly equicontinuous*: if $z, w \in \overline{\Omega}_m$ then $|f(z) - f(w)| \leq 2m|z - w|$. Let us take a sequence of functions $f_n$ in $\mathscr{F}$ such that

$$\lim_{n\to\infty} f_n'(z_0) \;=\; \sup_{f\in\mathscr{F}} f'(z_0).$$

By the Ascoli–Arzelá theorem (see below) there is a subsequence that converges uniformly on $\Omega_1$. A further subsequence converges uniformly on $\Omega_2$, and a standard argument gives us a subsequence that converges uniformly on each $\Omega_m$. The limit $f$ is a holomorphic function that maps $\Omega$ into $\mathbb{D}$ whose derivative $f'(z_0)$ has the desired maximum property. We need to show that $f$ is bijective. Suppose that $a_1$ and $a_2$ are two distinct points of $\Omega$ and $f(a_1) = f(a_2) = b$. Choose disjoint circles $C_j$ around $a_j$ with the property that $f(z) - b \neq 0$ on $C_1 \cup C_2$, and $f(z) - b$ has a single zero, counting multiplicity, inside each of the circles $C_j$. Let $r > 0$ be a lower bound for $|f(z) - b|$ on the union of the $C_j$, and let $n$ be so large that $|f_n(z) - f(z)| < r$ on this union. By Rouché's theorem, Theorem 1.3.7, $f_n(z) - b$ also has a zero inside each circle, contradicting the assumption that the $f_n$ are bijective. Thus $f$ is bijective. □

To complete the argument, we need to prove the Ascoli–Arzelá theorem.

**Theorem 5.3.1.** *(Ascoli–Arzelá) Suppose that $\{f_n\}$ is a uniformly bounded, uniformly equicontinuous family of real or complex-valued functions defined on a compact set $C$. Then there is a subsequence that converges uniformly.*

*Proof:* Assume first, for purpose of visualization, that the domain is the interval $[0,1]$, and the functions take values in $[0,1]$. Thus the graph of each function lies in the unit square. Partition the range into intervals of length $1/4$. Using the assumption of uniform equicontinuity, we can partition the domain into closed intervals such that if $a, b$ lie in one such interval, then $|f_n(a) - f_n(b)| < 1/4$, for each $n$. This partitions the square into closed rectangles, each with height $1/4$. Let $P_1(f_n)$ be the union of the rectangles that are intersected by the graph of $f_n$. This polygon intersects each vertical line over the interior of an interval of the partition of the domain in a segment of length at most $1/2$,

There are only finitely many distinct such polygons $P_1(f_n)$, so there must be such a polygon $P_1$ that contains the graphs of each term of a subsequence of the original sequence. Any two elements $f_m, f_n$ of this subsequence have $|f_n(x) - f_m(x)| \leq 1/2$ for every $x$.

Now partition the range into subintervals of length 1/8. Choose a refinement of the previous partition of the domain so that the variation over each new subinterval is $< 1/8$. This gives a partition of $P_1$ into closed rectangles of height $1/8$. There is a further subsequence whose graphs lie in a particular polygon $P_2 \subset P_1$ that has vertical height $\leq 1/4$ over each open subinterval of the domain. See Figure 5.1 for an illustration of this stage of the construction.

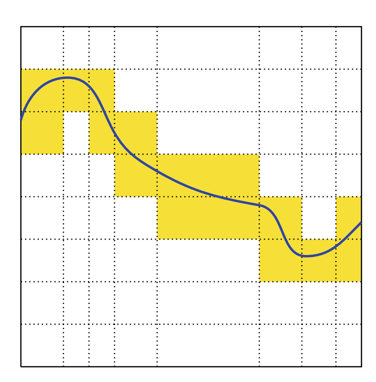


**Fig. 5.1** The polygon $P_2(f)$

Continuing in this way, taking terms with increasing subscripts from the successive subsequences, we get a subsequence that converges uniformly; see Exercise 2. This argument generalizes easily, though less pictorially, to the general case. □

## 5.4 Boundary behavior of conformal maps

By itself, the Riemann mapping theorem tells us little about the behavior of the map near the boundary. The one simple fact is the following.

**Lemma 5.4.1.** *If $f : \Omega \to \mathbb{D}$ is a holomorphic bijection from a bounded domain $\Omega$ onto the unit disk $\mathbb{D}$, then $|f(z)| \to 1$ as $z \to \partial\Omega$, the boundary of $\Omega$.*

*Proof:* Given $\varepsilon > 0$, let $C_\varepsilon$ be the inverse image of $\{z : |z| \leq 1-\varepsilon\}$. This is a compact subset of $\Omega$, so it lies at some positive distance $\delta$ from $\partial\Omega$. □

**Theorem 5.4.2. (Schwarz reflection principle)** *With $f$ as in the preceding proposition, suppose that the boundary $\partial\Omega$ contains $I$, an open straight line segment or an open circular arc, and suppose that points of $\Omega$ approach $I$ from only one side. Then $f$ extends to be holomorphic in a neighborhood of $I$.*

*Proof:* Since we can use a linear fractional transformation to straighten a circular arc, it is enough to consider the case of a line segment $I$. Suppose $z_0 \in I$. Choose $r$ small enough that the intersection of $\Omega$ with the disk $D_r(z_0)$ is a half-disk $D_+$, and the diameter of $D_+$ is contained in $I$. We may also suppose that $f$ has no zeros in this half-disk. Choose a branch of $\log f$ defined on $D_+$. By Lemma 5.4.1,

$$\mathrm{Re}\,[\log f(z)] \;=\; \log|f(z)| \;\to 0$$

as $z$ approaches the bounding diameter. By Theorem 4.3.2, $g = \log f$ has a holomorphic extension to the whole disk, so the same is true of $f = \exp g$. □

## 5.5 Mapping polygons: the Schwarz–Christoffel formula

The other way to look at the Riemann mapping theorem is that, given a simply connected proper open subset $\Omega \subset \mathbb{C}$, there is a holomorphic bijection from the unit disk $\mathbb{D}$ onto $\Omega$. In view of the fact that $\mathbb{D}$ and the half-plane $\mathbb{C}_+$ can be mapped to each other by linear fractional transformations, one could as well consider the question of mapping $\mathbb{C}_+$ to $\Omega$. For the simplest cases, one can find the explicit *form* of such a mapping.

Suppose that $\Omega$ is a simply connected plane domain whose boundary is a polygon—a union of straight line segments—and that the portion of $\Omega$ near any such segment lies on one side of that segment. Let $f$ be a holomorphic bijection from $\Omega$ onto the upper half plane. We know from Theorem 5.4.2 that $f$ has a holomorphic continuation across each of the sides of the polygon. By analyzing the behavior of $f$ near a corner, we can determine the form of the inverse map $F : \mathbb{C}_+ \to \Omega$. Note that $G(z) = F(1/z)$ maps $\mathbb{R}$ to the image of a side of $P$, which is either a line segment or a circular arc. Therefore $G$ continues across the real axis near $z = 0$, so $F$ is holomorphic near $\infty$.

**Theorem 5.5.1.** (Schwarz–Christoffel) *Let the domain $P$ be a polygon in $\mathbb{C}$ with vertices $z_1, z_2, \ldots, z_n$ and sides that are straight line segments. Suppose $P$ lies to the left of each segment as oriented from $z_k$ to $z_{k+1}$, with $z_{n+1}$ defined to be $z_1$. Let $\alpha_k \pi$, $0 < \alpha_k < 2$ be the angle, at $z_k$, from the segment with endpoints $z_k$ and $z_{k+1}$, to the segment with endpoints $z_{k-1}$ and $z_k$; see Figure 5.2.*

*Suppose that $F$ is a conformal map from $\mathbb{C}_+$ onto $P$. Then $F$ has continuous extension to the closure $\mathbb{C}_+ \cup \mathbb{R}$. There are constants $A$ and $B$ such that*

$$F(w) \ = \ A + B \int_0^w \prod_{k=1}^n (z - a_k)^{\alpha_k - 1} \, dz, \qquad (5.5.1)$$

*where $F(a_k) = z_k$.*

**Remark.** The sum of the angles of a triangle is $\pi$, and it can be deduced from this fact that the sum of the interior angles of an $n$-sided polygon is $(n-2)\pi$. Thus

$$\alpha_1 + \alpha_1 + \cdots + \alpha_n \ = \ n - 2. \qquad (5.5.2)$$

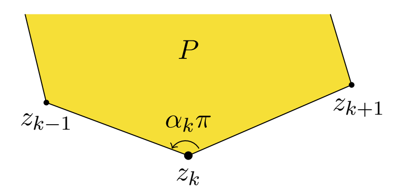



**Fig. 5.2** The angle at $z_k$

Therefore the integrand in (5.5.1) is $O(|z|^{-2})$ as $z \to \infty$, so the function defined by (5.5.1) is bounded.

Let $f : P \to \mathbb{C}_+$ be the inverse map. The proof of Theorem 5.5.1 begins with an examination of $f$ near a corner.

**Lemma 5.5.2.** *The map $f$ extends continuously to the closure of $P$. Let $a_k = f(z_k)$. The inverse function $F$ extends to a full neighborhood of $a_k$ in the closure of $\mathbb{C}_+$ and has the form*

$$F(w) = z_k + (w - a_k)^{\alpha_k} H_k(w), \tag{5.5.3}$$

*where $H_k$ is holomorphic and non-zero in a neighborhood of $a_k$.*

*Proof:* Choose a branch of $(z - z_k)^{1/\alpha_k}$ in the intersection of $P$ with a small disk centered at $z_k$. The choice of the power means that the image is a half-disk: the angle at $z_k$ has been changed to $\pi$. The map

$$g(\zeta) = f(z_k + \zeta^{\alpha_k})$$

maps this half-disk into $\mathbb{C}_+$. By Theorem 5.4.2, $g$ continues holomorphically across to the full disk. In particular, this implies that $f$ continues across a neighborhood of $z_k$. Since $f$ is injective, it follows that $g'(0) \neq 0$. Therefore there is a holomorphic inverse $G$, defined near $a_k$:

$$w = f(z_k + G(w)^{\alpha_k}),$$

or

$$F(w) = z_k + (w - a_k)^{\alpha_k} \left[ \frac{G(w)}{w - a_k} \right]^{\alpha_k}.$$

Now $G(a_k) = 0$, so $h_k(w) = G(w)/(w - a_k)$ is holomorphic and non-zero near $a_k$. Then $H_k = h_k^{\alpha_k}$ is holomorphic and non-zero near $a_k$. □

*Proof of Theorem* 5.5.1. It follows from Lemma 5.5.2 that, in a neighborhood of $a_k$,

$$F'(w) = [(w - a_k)^{\alpha_k} H_k(w)]' = (w - a_k)^{\alpha_k - 1}[\alpha_k H_k(w) + (w - a_k) H_k'(w)],$$

where $H_k$ is holomorphic near $a_k$. Therefore the function $H$ defined by

$$H(w) = \prod_{k=1}^{n} (w - a_k)^{1 - \alpha_k} F'(w) \tag{5.5.4}$$

is holomorphic, with no zeros, in a neighborhood of the closed half-plane $\mathbb{C}_+ \cup \mathbb{R}$. We claim that $\arg H$ is constant. Consider the argument of the restriction of $H(w)$ to $\mathbb{R}$. The argument is a continuous function of $w \in \mathbb{R}$. But the argument is also piecewise constant: the argument of each factor in the product (5.5.4) is constant between the points $a_k$ Therefore $\arg H(w)$ is constant for $w \in \mathbb{R}$. Note that $\arg H$, being the

imaginary part of $\log H$, is harmonic. We would like to invoke the maximum principle and conclude that $\arg H$ is constant on $\mathbb{C}_+$; however $\mathbb{C}_+$ is not bounded, so some additional reasoning is needed. (Consider the function $\mathrm{Im}\, z$, which vanishes on $\mathbb{R}$ but not on $\mathbb{C}_+$.)

We noted earlier that $F$ is holomorphic near $\infty$, and it follows that $F'(z) \sim cz^{-2}$ as $z \to \infty$. Combining this observation with (5.5.2), we see that $H \sim c$, for some constant $c$, as $z \to \infty$. Therefore $\arg z = \arg c$ on $\mathbb{R}$ and $\arg z \to \arg c$ as $z \to \infty$. A simple extension of the maximum principle shows that $\arg H \equiv \arg c$ on $\mathbb{C}_+$: Exercise 4 of Chapter 4. A non-constant holomorphic function cannot take all its values in a line, so $H$ is constant. The Schwarz–Christoffel formula (5.5.1) follows by solving (5.5.4) for $F'$ and integrating. □

The representation (5.5.1) is only unique up to linear fractional transformations that map $\mathbb{C}_+$ to itself, so we may choose three of the $a_k$ arbitrarily. In fact we can send one of the $a_k$ to infinity. Suppose $a_n \neq 0$ and replace the factor $(w - a_n)^{\alpha_n - 1}$ with

$$\left(\frac{w}{a_n} - 1\right)^{\alpha_n - 1}$$

and let $a_n \to \pm\infty$. Then (5.5.1) reduces to the form

$$F(w) = A + B \int_0^w \prod_{k=1}^{n-1} (z - a_k)^{\alpha_k - 1}\, dz, \tag{5.5.5}$$

(with different constants $A, B$).

Let us count real parameters. It takes $2n$ real parameters to specify the vertices $z_k$. In the formula (5.5.1) there are $n-1$ independent angles, $n-3$ independent points $w_k$, and four real parameters in $A$ and $B$. Thus the parameter count is consistent. Nevertheless, although every such conformal map of $\mathbb{C}_+$ is given by a formula (5.5.1), not every such formula produces a polygon with the desired properties. Assuming that (5.5.2) holds, the image of $\mathbb{R}$ under (5.5.1) will be a closed polygonal line with the specified angles, but this line may intersect itself in one or more places.

## 5.6 Triangles and rectangles

The simplest case of the formula in the form (5.5.5) gives the *Schwarz triangle functions*:

$$F(u) = \int_0^u w^{\alpha_1 - 1}(w-1)^{\alpha_2 - 1}\, dw. \tag{5.6.1}$$

As to rectangles, we have our choice of two standard canonical forms:

$$F(u) = \int_0^u \frac{dw}{\sqrt{(1-w^2)(1-k^2w^2)}}, \qquad k > 0,\ k^2 \neq 1, \tag{5.6.2}$$

and

$$F(u) = \int_0^u \frac{dw}{\sqrt{w(w-1)(w-\rho)}}, \qquad \rho > 1. \tag{5.6.3}$$

The inverses of these functions are the principal subjects of Chapter 15 and Chapter 16, respectively.

For either the triangle functions (5.6.1), or the rectangle functions (5.6.2) or (5.6.3), if we specify a side of the image, then the map $F$ extends to a map of $\mathbb{C}_- = \{z : \mathrm{Im}\, z < 0\}$ onto the reflection of the triangle or rectangle through that side, so as to map the plane minus the complement of an interval holomorphically onto the doubled triangle or rectangle.

For the rectangle as given by (5.6.2) or (5.6.3), this process can be iterated indefinitely, giving a doubly periodic map of the plane to itself (with poles at the corners of the various images of the original rectangle).

For the triangle maps, the process may or may not proceed indefinitely. For example, if we fix a vertex of the triangle and reflect around successive sides having that vertex, say in the clockwise direction, we eventually add a triangle that intersects the original (open) triangle. Thus for consistency it is necessary that the new triangle must coincide with the original triangle. Moreover the number of reflections involved must be even, because otherwise the extended map will take the new triangle to $\mathbb{C}_-$. Checking the angles at the fixed vertex, we see that the corresponding $\alpha_k$ must be the reciprocal of an even integer. For consistency, this must be true of each vertex. This means that there are only four possible cases; see Exercise 7.

## 5.7 Univalent functions

The term *univalent* means single-valued, i.e. injective, not taking the same value twice. In complex function theory the term is primarily used for holomorphic functions, specifically for single-valued holomorphic functions defined on the disk $\mathbb{D}$. If $f$ is such a function and $f(\mathbb{D}) = \Omega$, then the inverse $f^{-1} = g : \Omega \to \mathbb{D}$ is one of the maps whose existence is guaranteed by the Riemann mapping theorem. By composing with a (unique) affine map $z \to az + b$ we may impose additional conditions

$$f(0) = 0, \qquad f'(0) = 1. \tag{5.7.1}$$

Then the Maclaurin expansion (Taylor expansion at the origin) is

$$f(z) = z + a_2 z^2 + a_3 z^3 + a_4 z^4 + \ldots, \quad |z| < 1. \tag{5.7.2}$$

The set of univalent maps $f$ that are defined on $\mathbb{D}$ and satisfy (5.7.1) is denoted $S$. The $S$ stands for *schlicht*, German for "simple." Functions that belong to $S$ are often called *schlicht functions*.

The uniqueness part of the Riemann mapping theorem implies that if $f$ and $g$ are in $S$, and $f(\mathbb{D}) = g(\mathbb{D})$, then $f = g$: Exercise 14. Therefore it is natural to examine

relations between properties of the image $\Omega = f(\mathbb{D})$ and properties of the coefficients in the expansion (5.7.2).

A particularly important example comes about as follows: the linear fractional transformation

$$f(z) = \frac{1+z}{1-z}$$

maps $\mathbb{D}$ to the right half plane $\{z : \operatorname{Re} z > 0\}$. Therefore $f(z)^2$ maps $\mathbb{D}$ to the complement of the half-line $\{x : x \le 0\}$. We can adjust this to get a function in $S$ by a translation and dilation. The result is the *Koebe function*

$$\begin{aligned} K(z) = \frac{1}{4}[f(z)^2 - f(0)^2] &= \frac{z}{(1-z)^2} \\ &= z + 2z^2 + 3z^3 + 4z^4 + \dots . \end{aligned} \tag{5.7.3}$$

The image of $\mathbb{D}$ under $K$ is the complement of the half-line $\{x : x \le -1/4\}$. More generally, given $\theta \in \mathbb{R}$, define

$$K_\theta(z) = e^{-i\theta} K(e^{i\theta} z) = z + \sum_{n=2}^{\infty} a_n z^n, \quad a_n = n\, e^{i(n-1)\theta}.$$

Then $|a_n| = n$, and the complement of the image $K_\theta(\mathbb{D})$ is a rotation of the half-line $\{x : x \le -1/4\}$.

The Koebe functions $K_\theta$ have a number of extremal properties among the functions in $S$. Bieberbach proved in 1916 that for each $f \in S$, the term $a_2$ in the expansion (5.7.2) satisfies the estimate $|a_2| \le 2$, with equality only if $f$ is one of the functions $K_\theta$; [23]. Bieberbach conjectured that $|a_n| \le n$ for all $n$ and all $f \in S$. Throughout much of the 20th century various special cases of Bieberbach's conjecture were proved: see Duren [40]. The full conjecture was finally proved by de Branges in 1984 [34]. (De Brange's original manuscript ran to 385 typed pages. The proof was then considerably simplified by de Branges and others.) In this section we present Bieberbach's proof for the case $n = 2$.

If $f$ belongs to $S$, then the function

$$g(z) = \frac{1}{f(1/z)} = z\left[1 + a_2 z^{-1} + a_3 z^{-2} + \dots\right]^{-1}, \quad |z| > 1, \tag{5.7.4}$$

is univalent and has a simple pole at $\infty$. Let us consider functions of this type:

$$h(z) = z + b_0 + b_1 z^{-1} + b_2 z^{-2} + b_3 z^{-3} + \dots, \quad |z| > 1. \tag{5.7.5}$$

A key result is due to Gronwall [54]:

**Theorem 5.7.1.** (Area Theorem) *If a function $h$ given by the formula (5.7.5) is univalent, then*

$$\sum_{n=1}^{\infty} n|b_n|^2 \le 1. \tag{5.7.6}$$

*Proof:* For $r > 1$, let $E_r$ be the complement of the image $\{h(z) : |z| \geq r\}$, and let

$$C_r = \{h(z) : |z| = r\}.$$

Note that $C_r$ is a simple closed curve that encloses $E_r$. By (1.1.8), the area of $E_r$ is

$$\begin{aligned} A_r &= \frac{1}{2i}\int_{|z|=r} \overline{h(z)}h'(z)\,dz \\ &= \frac{1}{2}\int_0^{2\pi}\left[re^{-i\theta} + \sum_{n=0}^{\infty}\bar{b}_n r^{-n}e^{in\theta}\right]\left[re^{i\theta} - \sum_{m=1}^{\infty} m b_m r^{-m}e^{-im\theta}\right]d\theta. \end{aligned}$$

The series converge uniformly, so we may take the product and integrate term-by-term. Since

$$\frac{1}{2}\int_0^{2\pi} e^{ip\theta}\,d\theta = \begin{cases} \pi & \text{if } p = 0, \\ 0 & \text{if } p = \pm 1, \pm 2, \ldots \,. \end{cases}$$

it follows that

$$A_r = \pi\left[r^2 - \sum_{n=1}^{\infty} n|b_n|^2 r^{-2n}\right].$$

Letting $r$ decrease to 1, the limit of the left side is the outer measure $m^*(E)$ of the complement $E$ of the image of the map $h$. Therefore

$$0 \leq m^*(E) = \pi\left[1 - \sum_{n=1}^{\infty} n|b_n|^2\right]. \qquad \square$$

In particular, equality holds in (5.7.6) if and only if the complement of the image of $h$ has measure zero. If $h = g$ has the form (5.7.4), where $f$ belongs to $S$, then this is equivalent to saying that the complement of the image of $f$ has measure zero.

**Corollary 5.7.2.** *If $h$ given by (5.7.5) is univalent then for each $n$, $|b_n|^2 \leq 1/n$. Moreover $|b_1| = 1$ if and only if the complement of the image of $h$ is a line segment of length 4.*

*Proof:* The first assertion is immediate. If $|b_1| = 1$, then $b_n = 0$ for $n > 1$:

$$h(z) = z + b_0 + \frac{b_1}{z}. \tag{5.7.7}$$

Suppose first that $b_1 = 1$. Then the product of the two solutions to $h(z) = w$ is $\equiv 1$, so there is a unique solution $z$ in the complement of $\mathbb{D}$ unless both solutions have modulus 1. This occurs if and only if $w$ is in the line segment with endpoints $b_0 - 2$, $b_0 + 2$; Exercise 20. Replacing $z$ with $\omega z$, where $\omega^2 = b_1$, reduces the problem to the case with $b_1$ replaced by 1 and $b_0$ replaced by $\omega^{-1}b_0$ so the exceptional set is the line segment with endpoints $\omega^{-1}b_0 - 2$, $\omega^{-1}b_0 + 2$. $\square$

The remaining ingredient in Bieberbach's proof is the square-root transformation. Suppose $f$ belongs to $S$. Then

$$f(z^2) \;=\; z^2\left[1+\sum_{n=1}^{\infty} a_{n+1}z^{2n}\right], \quad |z|<1.$$

By assumption, $f$ is univalent, so the term in brackets is never 0. Therefore we may choose a branch of the square root that is 1 at $z=0$ and define

$$f_2(z) \;\equiv\; f(z^2)^{1/2} \;=\; z\left[1+\frac{a_2}{2}z^2+\dots\right], \quad |z|<1.$$

This is easily seen to be single-valued, so it belongs to $S$.

**Theorem 5.7.3.** (Bieberbach) *If $f$ belongs to $S$ and has the expansion (5.7.2), then $|a_2|\le 2$. The equality is strict unless $f=K_\theta$ for some $\theta$.*

*Proof:* Let

$$g(z) \;=\; \frac{1}{f_2(1/z)} \;=\; \frac{1}{f(1/z^2)^{1/2}} \;=\; z-\frac{a_2}{2}z^{-1}+\dots\,. \tag{5.7.8}$$

By Corollary 5.7.2, $|a_2|\le 2$. Equality implies that

$$g(z) \;=\; z-\frac{e^{i\theta}}{z}, \tag{5.7.9}$$

for some $\theta\in\mathbb{R}$. It follows that $f=K_\theta$. □

## Exercises

1. Prove the estimate (5.3.1).
2. Prove the statement in the proof of the Ascoli–Arzelá theorem that a subsequence can be chosen so that it converges on each of the $\Omega_m$. (Hint: the construction is known as the "diagonal process.")
3. Suppose that the simply connected domain $\Omega$, a proper subset of $\mathbb{C}$, is invariant under rotation by an angle $2\pi/k$, $k$ an integer $\ge 2$, i.e. $z\in\Omega$ implies $\omega z\in\Omega$, where $\omega=\exp(2\pi i/k)$. Show that 0 belongs to $\Omega$. Show that the proof of the Riemann mapping theorem can be adapted to show that there is a conformal map $f:\Omega\to\mathbb{D}$ such that $f(0)=0$ and such that $f$ commutes with rotation by $2\pi/k$: $f(\omega z)=\omega f(z)$.
4. Find the image $f(\mathbb{D})$ of the unit disk $\mathbb{D}$ for the functions:

   (a) $f(z)=-\log(1-z)$. (b) $f(z)=\tan^{-1}z$.
5. Show that the integral (5.5.1) has finite limits as $w\to\pm\infty$.
6. Show that the integral (5.5.5) has finite limits as $w\to\pm\infty$.

7. Show that the inverse of the triangle map (5.6.1) extends to a meromorphic map on the whole plane if and only if its angles are one of the three triples
$$\left\{\frac{\pi}{3},\frac{\pi}{3},\frac{\pi}{3}\right\}, \qquad \left\{\frac{\pi}{2},\frac{\pi}{3},\frac{\pi}{6}\right\}, \qquad \left\{\frac{\pi}{2},\frac{\pi}{4},\frac{\pi}{4}\right\}.$$
Show that in each case, the resulting function is doubly periodic. Note that in the second case, there are two non-congruent triangles with this triple of angles.
8. Show that the image of $\mathbb{D}$ under the map
$$f(z) = \int_0^z (1-s^n)^{-2/n}\, ds, \quad n \geq 3,$$
is a regular $n$-sided polygon centered at the origin.
9. Suppose
$$f'(z) = -2i\left(z - \frac{1}{z}\right).$$
Show that the constant of integration can be chosen so that $f$ maps $\mathbb{C}_+$ onto $\mathbb{C}\setminus J$, where $J$ consists of the two half-lines $\operatorname{Re} w = \pm\pi$, $\operatorname{Im} w \geq 0$.
10. Suppose
$$f'(z) = \frac{z}{\sqrt{z^2-a^2}}.$$
Show that with the correct choices of $a$, the square root, and the constant of integration, $f$ maps $\mathbb{C}_+$ onto $\mathbb{C}_+ \setminus J$, where $J$ is the interval from 0 to $ia$.
11. (a) Find a conformal map from the half-disk $\{z : |z| < 1, \operatorname{Im} z > 0\}$ onto $\mathbb{C}_+$.
(b) Find a conformal map from the half-disk onto the full disk $\mathbb{D}$.
12. Find the image under $f(z) = \frac{1}{2}(z + \frac{1}{z})$ of the following:
(a) $\{z : |z| > 1\}$;
(b) $C_+$;
(c) $\{z : 1 < |z|, \operatorname{Im} z > 0\}$.
(d) $\{z : |z| = r\}, \quad 0 < r \neq 1$.
13. Given $a > b > 0$, $a > 1$, find a conformal map from the domain bounded by the ellipses
$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \qquad \frac{x^2}{a^2+k^2} + \frac{y^2}{b^2+k^2} = 1;$$
onto a domain bounded by two circles $x^2+y^2 = r^2$, $x^2+y^2 = R^2$. (Hint: see (d) of Exercise 12.)
14. Prove that if $f$ and $g$ belong to the set $S$ of normalized univalent functions defined on $\mathbb{D}$, and $f(\mathbb{D}) = g(\mathbb{D})$, then $f = g$.
15. Show that each of the following functions belongs to the set $S$, and describe its range $f(\mathbb{D})$:

$$
\begin{aligned}
&\text{(a)} \quad f(z) = \frac{z}{1-z};\\
&\text{(b)} \quad f(z) = \frac{z}{1-z^2};\\
&\text{(c)} \quad f(z) = \frac{1}{2}\log\frac{1+z}{1-z};\\
&\text{(d)} \quad f(z) = z - \frac{1}{2}z^2.
\end{aligned}
$$

16. Suppose that $\sum_{n=2}^{\infty} n|a_n| \leq 1$. Prove that

$$f(z) = z + a_2 z^2 + a_3 z^3 + \dots$$

belongs to $S$.
17. Show that $h$ given by (5.7.7) is univalent from $\{z : |z| > 1\}$ onto the complement of a line segment of length 4.
18. Show that (5.7.8) and (5.7.9) imply that $f = K_\theta$.
19. Show that if $f \in S$ and $a$ is not in $f(\mathbb{D})$, then the function

$$g(z) = \frac{af(z)}{a - f(z)} = z + \left(a_2 + \frac{1}{a}\right)z^2 + \dots$$

belongs to $S$.
20. Prove that $z + b_0/z + 1 = w$ has a unique solution $z$ with modulus $|z| > 1$ if and only if $w$ is not in the line segment $[b_0 - 2, b_0 + 2]$.
21. Use Exercise 19 and Bieberbach's theorem to prove the theorem known as the *Koebe one-quarter theorem*: if $f \in S$ then $f(\mathbb{D})$ includes the disk $D_{1/4}(0) = \{z : |z| < 1/4\}$. Moreover $f(\mathbb{D})$ includes a larger disk, unless $f = K_\theta$ for some $\theta$. (Actually Koebe proved that there was a $\delta > 0$ such that $D_\delta(0)$ is included in each $f(\mathbb{D})$, and conjectured that the maximal such radius was 1/4. Bieberbach proved Koebe's conjecture.)

## Remarks and further reading

Riemann's version of Theorem 5.2.1 used some arguments that were only justified (in part) later. For a succinct discussion of the history of Theorem 5.2.1, see Section 17.1 of Hille [64]. For a detailed account see Gray [53]. Ahlfors [5] contains an efficient treatment of mappings of multiply connected domains. A classic account of conformal mapping is that by Nehari [104]. For more, see Bell [17] and Kythe [82].

Chapter 6 covers additional results related to the material in Sections 5.5 and 5.6 on mappings to polygons.

As noted above, much more information about univalent functions can be found in Duren. For the interaction with deformation theory, see Lehto [87].

An important generalization of a conformal map is a *quasiconformal map*. As we have seen, a conformal map $f$, defined on a domain in $\mathbb{C}$, that preserves orientation satisfies the equation $\bar{\partial} f = 0$. (For the notation here, see Section 1.1.) A quasiconformal map satisfies the *Beltrami equation* $\bar{\partial} f = \mu \partial f$, with $\sup |\mu(z)| < 1$. Quasiconformal maps are a fundamental tool in the study of deformations of certain geometric structures; see Ahlfors and Bers [6]. Bers [22] provides a readable and comprehensive overview. For more details, see Lehto [87].

# Chapter 6
# The Schwarzian derivative



The Schwarzian derivative of a function $f$ is a rational function of the derivatives of $f$ to order 3. In fact it can be expressed in terms of the logarithmic derivative $f''/f'$ of $f'$. Here we show that the Schwarzian derivative is a natural object: a measure of the "curvature" of $f$, the pointwise deviation from a best approximation of $f$ by a linear fractional transformation.

The Schwarzian derivative was introduced by H. Schwarz in his study of conformal maps from the disk or half plane to a polygon—including polygons whose sides may be arcs of circles rather than straight line segments. The global extension of such a map, by continued reflection across boundaries, is not single-valued, in general. Nevertheless its Schwarzian derivative is single-valued. A consequence is that the extended map may be realized as the quotient of two single-valued functions that are solutions of a second-order linear differential equation of special (Fuchsian) type. In the case of a triangle or a regular curvilinear polygon, the map is a quotient of two hypergeometric functions.

## 6.1 The Schwarzian derivative as measure of curvature

Suppose that $\gamma : [a,b] \to \mathbb{R}^2$ is a simple curve in the plane. The curvature of $\gamma$ at a point $\gamma(t_0)$ is a measure of the deviation of the image from the tangent line at $\gamma(t_0)$. Equivalently, the curvature is a measure of the deviation of the function $\gamma$ from the affine transformation $g(t) = \gamma(t) + \gamma'(t)(t - t_0)$. If we consider holomorphic functions, and consider the complex plane as part of the Riemann sphere $\mathbb{S}$, then affine transformations play no special role—but linear fractional transformations do play a special role.

Suppose that $f$ is holomorphic in a neighborhood of a point $z \in \mathbb{C}$, and $f'(z) \neq 0$. Given $\varepsilon$ sufficiently close to 0, the points $z, z+\varepsilon, z+2\varepsilon, z+3\varepsilon$ all lie in the domain of $f$. Consider the cross-ratio

$$[f(z), f(z+\varepsilon), f(z+2\varepsilon), f(z+3\varepsilon)]. \tag{6.1.1}$$



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_6

This is constant if $f$ is a linear fractional transformation. In fact the cross-ratio is invariant under linear fractional transformation. Therefore if $f$ itself is a linear fractional transformation, we may compose with $g(z) = \varepsilon^{-1}[f^{-1}(z) - z]$ and conclude that (6.1.1) is equal to the cross-ratio $[0,1,2,3]$, which is $= -1/3$.

It follows that the deviation of (6.1.1) from $-1/3$, for small $\varepsilon$, is a measure of the deviation of $f$, at $z$, from being a linear fractional transformation. We shall see that (6.1.1) is $-1/3 + O(\varepsilon^2)$. Therefore the coefficient of $\varepsilon^2$ in the expansion of (6.1.1) is a measure of the curvature of $f$ at $z$, in this sense.

To compute this coefficient, we take terms up to order $O(\varepsilon^3)$ in the Taylor expansion of $f$ at a given point $z$. The cross-ratio depends only on differences, so we may subtract $f(z)$ from each term of (6.1.1). The cross-ratio is homogeneous of degree zero in its arguments, so we may divide each term by $f'(z)\varepsilon$. This leads us to define

$$g_1 = \frac{f''(z)\varepsilon}{2f'(z)}, \qquad g_2 = \frac{f'''(z)\varepsilon^2}{6f'(z)}. \tag{6.1.2}$$

Then

$$\frac{f(z+k\varepsilon) - f(z)}{\varepsilon f'(z)} = k + k^2 g_1 + k^3 g_2 + O(\varepsilon^3).$$

Therefore

$$\begin{aligned}
&[f(z), f(z+\varepsilon), f(z+2\varepsilon), f(z+3\varepsilon)] \\
&\quad = [0,\ 1+g_1+g_2,\ 2+4g_1+8g_2,\ 3+9g_1+27g_2] + O(\varepsilon^3).
\end{aligned}$$

Taking into account the definition

$$[w_0, w_1, w_2, w_3] = \frac{(w_0 - w_1)(w_2 - w_3)}{(w_0 - w_3)(w_2 - w_1)} = -\frac{(w_1 - w_0)(w_3 - w_2)}{(w_3 - w_0)(w_2 - w_1)},$$

we have

$$\begin{aligned}
&-[0,\ 1+g_1+g_2,\ 2+4g_1+8g_2,\ 3+9g_1+27g_2] \\
&\quad = \frac{(1+g_1+g_2)(1+5g_1+19g_2)}{3(1+3g_1+9g_2)(1+3g_1+7g_2)} + O(\varepsilon^3) \\
&\quad = \frac{1}{3}\,\frac{1+6g_1+5g_1^2+20g_2}{1+6g_1+9g_1^2+16g_2} + O(\varepsilon^3) \\
&\quad = \frac{1}{3}\,\frac{1+(5g_1^2+20g_2)(1+6g_1)^{-1}}{1+(9g_1^2+16g_2)(1+6g_1)^{-1}} + O(\varepsilon^3) \\
&\quad = \frac{1}{3}\,\frac{1+5g_1^2+20g_2}{1+9g_1^2+16g_2} + O(\varepsilon^3) \\
&\quad = \frac{1}{3}(1+5g_1^2+20g_2)(1-9g_1^2-16g_2) + O(\varepsilon^3) \\
&\quad = \frac{1}{3}(1-4g_1^2+4g_2) + O(\varepsilon^3).
\end{aligned}$$

Plugging in the definitions (6.1.2),

$$\begin{aligned}
&[f(z), f(z+\varepsilon), f(z+2\varepsilon), f(z+3\varepsilon)] \\
&= \frac{1}{3}\left[1+\frac{2}{3}\frac{f'''(z)}{f'(z)}\varepsilon^2 - \left(\frac{f''(z)}{f(z)}\right)^2\varepsilon^2\right] + O(\varepsilon^3) \\
&= \frac{1}{3}+\frac{2}{9}\{f,z\}\,\varepsilon^2 + O(\varepsilon^3),
\end{aligned}$$

where $\{f,z\}$ is the *Schwarzian derivative* (or simply the *Schwarzian*)

$$\{f,z\} = \frac{f'''(z)}{f'(z)} - \frac{3}{2}\left(\frac{f''(z)}{f'(z)}\right)^2 = \left(\frac{f''(z)}{f'(z)}\right)' - \frac{1}{2}\left(\frac{f''(z)}{f'(z)}\right)^2. \tag{6.1.3}$$

**Remark.** The Schwarzian derivative of $f$ has at most a double pole at a pole of $f$. Therefore the Schwarzian of a meromorphic function is itself meromorphic.

## 6.2 Some properties of the Schwarzian

The Schwarzian has several important invariance properties.

**Proposition 6.2.1.** *The Schwarzian derivative has the properties:*

(a) *The Schwarzian of a linear fractional transformation is zero.*

(b) *If $f$ is meromorphic in a connected domain and $\{f,z\} \equiv 0$, then $f$ is a linear fractional transformation.*

(c) *If $f$ and $g$ are smooth functions, then*

$$\{f\circ g, z\} = \{f, g(z)\}\, g'(z)^2 + \{g,z\}.$$

(d) *If $f$ is a smooth function and $g$ is a linear fractional transformation, then*

$$\{g\circ f, z\} = \{f,z\}.$$

(e) *If $f_1$ and $f_2$ are meromorphic in a connected region and $\{f_1,z\} \equiv \{f_2,z\}$, then there is a linear fractional transformation $g$ such that $f_2 = g\circ f_1$.*

*Proof:* The proof can be reconstructed from the following suggestions. The details are left as Exercise 2.

(a) This follows from the fact that for a linear fractional transformation $f$, (6.1.1) is constant. It is also easily checked by direct calculation.

(b) Let $F = f''/f'$ and integrate the resulting equation (6.1.3) for $F$.

(c) Calculate.

(d) Use (c) and (a).

(e) It is enough to prove this locally. In a domain in which $f_1$ has an inverse $h$, let $g = f_2 \circ h$. Then $f_2 = g \circ f_1$. Use the assumption that $\{f_2, z\} = \{f_1, z\}$, together with (c) and (b), to get the result. □

## 6.3 The Schwarzian and curves

Next, consider a smooth curve $t \to \gamma(t)$ in $\mathbb{C}$, where the parameter $t$ can be taken to be either real or complex, and we assume that the derivative $\gamma'$ is not zero. (When the parameter is complex, it is more natural to think of $\gamma$ as a function rather than a curve.) Then $\gamma$ can be considered as a curve in the Riemann sphere $\mathbb{S}$, realized via an equivalence relation in $\mathbb{C}^2$; see Section 2.1. A *lift* of $\gamma$ to $\mathbb{C}^2$ consists of two curves $\varphi_1$, $\varphi_2$ in $\mathbb{C}$ such that $\gamma$ is represented as a quotient

$$\gamma = \varphi_1/\varphi_2. \tag{6.3.1}$$

**Theorem 6.3.1.** *Suppose that the map $\gamma(t) : \Omega \to \mathbb{C}$ is smooth, and $\gamma' \neq 0$. Suppose also that there is a single-valued branch of $\sqrt{\gamma'}$ defined on $\Omega$. Then there is a lift $\gamma = \varphi_1/\varphi_2$ such that $\varphi_1$ and $\varphi_2$ are linearly independent solutions of the equation*

$$\varphi''(t) + q(t)\varphi(t) = 0, \tag{6.3.2}$$

*where*

$$q(t) = \frac{1}{2}\{\gamma, t\}. \tag{6.3.3}$$

*Conversely, given a function $q$, let $\psi_1$ and $\psi_2$ be linearly independent solutions of (6.3.2). Then the quotient $\gamma_1 = \psi_1/\psi_2$ also satisfies (6.3.3). There is a linear fractional transformation $g$ such that $\gamma_1 = g \circ \gamma$.*

*Proof:* Differentiating (6.3.1) gives

$$\gamma' = \frac{\varphi_1'\varphi_2 - \varphi_1\varphi_2'}{\varphi_2^2}. \tag{6.3.4}$$

Let us impose the additional condition

$$1 \equiv \varphi_1'\varphi_2 - \varphi_1\varphi_2'. \tag{6.3.5}$$

Conditions (6.3.1) and (6.3.5) together imply that

$$\gamma' = \frac{1}{\varphi_2^2}. \tag{6.3.6}$$

Our assumption on the domain means that if we take $\varphi_2 = 1/\sqrt{\gamma'}$ and take $\varphi_1 = \gamma/\sqrt{\gamma'}$, then both (6.3.4) and (6.3.6) are satisfied. Differentiating (6.3.5) and

dividing by $\varphi_1 \varphi_2$ we find a proportionality

$$\frac{\varphi_j''}{\varphi_j} = -q(t), \quad j = 1, 2.$$

Starting from (6.3.6) we see that

$$\frac{\gamma''}{\gamma'} = -2\frac{\varphi_2'}{\varphi_2},$$

and (6.3.3) follows quickly. (Note that in this calculation, the constant 1 in (6.3.5) may be replaced by any non-zero constant.)

Conversely, suppose $\psi_1$, $\psi_2$ are linearly independent solutions of (6.3.2). Linear independence implies that the Wronskian $\psi_1 \psi_2' - \psi_1' \psi_2$ is not identically zero, and differentiation shows that it is constant. The previous calculation shows that $\gamma_1 = \psi_1/\psi_2$ is a solution of (6.3.3).

Since $\gamma_1$ and $\gamma$ have the same Schwarzian derivative, the last assertion follows from Proposition 6.2.1 (e). A more constructive derivation follows from the fact that the space of solutions of a second-order equation like (6.3.2) is two-dimensional, so there are constants $a, b, c, d$ with $ad - bc \neq 0$ such that

$$\psi_1 = a\varphi_1 + b\varphi_2, \quad \psi_2 = c\varphi_1 + d\varphi_2.$$

It follows readily that

$$\gamma_1 = \frac{\psi_1}{\psi_2} = \frac{a\gamma + b}{c\gamma + d}. \qquad \square$$

## 6.4 The Riemann mapping function and the Schwarzian

Schwarz developed these ideas to study curvilinear polygons in the complex plane. A *curvilinear polygon* $P$ is a connected domain in $\mathbb{C}$ whose boundary consists of finitely many vertices $z_1, z_2, \ldots, z_n, z_{n+1} = z_1$, each successive pair joined by either an arc of a circle or a segment of a straight line, and such that points of $P$ approach only one side of each of these arcs. Such regions arise, for example, as fundamental domains in the study of automorphic functions—see Chapter 17.

We assume that the vertices are numbered so that $P$ lies on the left of the boundary arc or segment from $z_{k-1}$ to $z_k$, and that the two tangents at $z_k$ meet at an interior angle $\pi\alpha_k$, $0 < \alpha_k < 2$, $\alpha_k \neq 1$. This is the curvilinear analogue of Figure 5.2. By the Riemann mapping theorem there is an invertible holomorphic map $g$ that maps such a domain $P$ onto the upper half-plane $\mathbb{C}_+$. By the Schwarz reflection principle $g$ extends holomorphically across each of the arcs or segments that make up the boundary of $P$. As in the case of an ordinary polygon (with straight line segments making up the boundary), the map $g$ also extends continuously at each vertex.

**Lemma 6.4.1.** *The mapping function $g$ extends continuously to the boundary of $P$.*

*Proof:* If the two sides that meet at $z_k$ are each straight line segments, then Lemma 5.5.2 applies. Otherwise, since the extensions of the two sides to two full circles, or to a line and a circle, are not tangent to each other at $z_k$, these extensions must meet at a second point $w_k$. A linear fractional transformation $G$ that takes $w_k$ to $\infty$ maps both sides to straight lines. Therefore Lemma 5.5.2 applies to the composition $g_1 = g \circ (G^{-1})$, and we can conclude that $g = g_1 \circ G$ also has the extension property. □

We propose now to study the inverse map $f : \mathbb{C}_+ \to P$. Let $a_k$ be the image of the vertex $z_k$ under (the extension of) $g$. We may assume that the $z_k$ are numbered so that $a_1 < a_2 < \cdots < a_n$. We know that $f$ can be continued across each of the open intervals $(a_k, a_{k+1})$, $k < n$ and across the pair of intervals $(a_n, \infty) \cup (-\infty, a_1)$ which, for convenience, we denote $(a_n, a_{n+1})$. Each continuation maps the lower half plane to an image $P_k$ of $P$ "reflected" across the corresponding side with vertices $z_k$, $z_{k+1}$. Having crossed one of the intervals from $\mathbb{C}_+$ to $\mathbb{C}_-$, one can cross back through a different interval, and the new function on $\mathbb{C}_+$ maps to an image of $P_k$ reflected through a side of $P_k$. This process can be continued indefinitely, resulting in a (generally) multi-valued holomorphic extension of $f$, with singular points at the $a_k$.

Let us pause to look more closely at the reflection process. The two simple models are the unit circle under the reflection $z \to 1/\bar{z}$ and the real line under the reflection $z \to \bar{z}$. In general, for a circle with center $a$ and radius $r$, a point $z$ and its reflection $z'$ are related by

$$(z' - a)(\bar{z} - \bar{a}) \;=\; r^2.$$

For a line $L$ passing through two points $a_1$, $a_2$, the linear fractional transformation

$$G(z) \;=\; \frac{z - a_1}{a_2 - a_1}$$

takes $L$ to $\mathbb{R}$, so a point $z$ and its reflection are related by

$$0 \;=\; G(z') - \overline{G(z)} \;=\; \frac{z' - a_1}{a_2 - a_1} - \frac{\bar{z} - \bar{a}_1}{\bar{a}_2 - \bar{a}_1}.$$

In each case the reflected point has the form $z' = \overline{H(z)}$, where $H$ is a linear fractional transformation. Note that the composition of two such reflections is a linear fractional transformation.

It follows from this discussion that, in the process of extending the mapping function $f$, two different determinations of the values in $\mathbb{C}_+$ come from an even number of reflections through sides of $P$. Thus these two determinations are related by a linear fractional transformation. The same is true for different determinations of values in $\mathbb{C}_-$.

After these preliminaries, we are ready to discuss the Schwarzian derivative of the extension $f_e$ of $f$.

**Theorem 6.4.2.** (Schwarz) *Let $f_e$ be the extension of the conformal map $f : \mathbb{C}_+ \to P$, where $P$ is a curvilinear polygon with vertices $\{z_k = f(a_k)\}_{k=1}^n$ and interior angles $\{\pi\alpha_k\}_{k=1}^n$. The Schwarzian derivative is the single-valued rational function*

$$\{f_e, w\} \;=\; \frac{1}{2}\sum_{k=1}^{n}\left[\frac{1-\alpha_k^2}{(w-a_k)^2} + \frac{2b_k}{w-a_k}\right], \tag{6.4.1}$$

*where the $b_k$ are certain real numbers, and*

$$0 \;=\; \sum_{k=1}^{n} b_k \;=\; \sum_{k=1}^{n}\left(2b_k a_k + 1 - \alpha_k^2\right) \;=\; \sum_{k=1}^{n}\left[b_k a_k^2 + (1-\alpha_k^2)\,a_k\right]. \tag{6.4.2}$$

*Proof:* The original map $f$ is injective on $\mathbb{C}_+$, continuous up to the boundary $\mathbb{R}$, and differentiable except at the points $a_k$. Therefore $f'$ has no zeros here. The extension of $f$ across each of the intervals $(a_k, a_{k+1})$ is again injective, so the derivative of the extension also has no zeros. This remains true for every subsequent extension of $f$. Therefore $\{f_e, w\}$ is defined on the full extension. Since the Schwarzian is invariant under linear fractional transformations, it follows that the various extensions of $f$ that lie over any point $z$ other than the points $\{a_k\}$ have the same Schwarzian. Thus $\{f_e, w\}$ is holomorphic and single-valued in the plane minus the points $\{a_k\}$.

Given $k$, let $G$ be the linear fractional transformation used in the proof of Lemma 6.4.1. Using invariance once again, near $a_k$ we may replace $f_e$ by $f_k = G \circ f_e$. This mapping takes a neighborhood of $f(a_k)$ in $\mathbb{R}$ onto the two straight line segments that are images under $G$ of the sides of $P$ that meet at $z_k$. Since $G$ is conformal, the internal angle is again $\alpha_k\pi$. As shown in Chapter 5, near $a_k$

$$f_k(w) \;=\; z_k + (w-a_k)^{\alpha_k} h_k(w), \tag{6.4.3}$$

where $h_k$ is holomorphic near $a_k$. Therefore

$$f_k'(w) \;=\; (w-a_k)^{\alpha_k - 1}[\alpha_k h_k(w) + (w-a_k)h_k'(w)]$$

and

$$\frac{f_k''(w)}{f_k'(w)} \;=\; \frac{\alpha_k - 1}{w-a_k} + \widetilde{h}_k(w),$$

where $\widetilde{h}_k$ is holomorphic near $a_k$. It follows that, near $a_k$,

$$\begin{aligned}
\{f_e, w\} &= \{f_k, w\} \\
&= \frac{1-\alpha_k}{(w-a_k)^2} - \frac{1}{2}\frac{(\alpha_k-1)^2}{(w-a_k)^2} - \widetilde{h}_k(w)\frac{\alpha_k-1}{w-a_k} - \frac{1}{2}\widetilde{h}_k^2(w) + \widetilde{h}_k'(w) \\
&= \frac{1-\alpha_k^2}{2(w-a_k)^2} - \widetilde{h}_k(w)\frac{\alpha_k-1}{w-a_k} - \frac{1}{2}\widetilde{h}_k^2(w) + \widetilde{h}_k'(w).
\end{aligned}$$

Thus the singularity of $\{f,w\}$ at $a_k$ is a double pole of the form in (6.4.1).

Now $f$ is real on each interval $(a_k, a_{k+1})$. This follows once again from invariance, since the corresponding side of $P$ can be mapped to a real interval by a linear fractional transformation $H$, and the corresponding map $H \circ f$ is real-valued on this interval. Therefore the constants $b_k$ that occur as residues at the $a_k$ are real.

To show that $\{f_e, w\}$ is determined by its singularities, we note that $f_e$ is holomorphic at $\infty$: $f(\infty)$ lies in the image of one of the sides of $P$. This implies that $\{f_e, w\}w^4$ is holomorphic at $\infty$; see Exercise 3. Therefore each side of (6.4.1) approaches zero at $\infty$. The difference between the left and right sides of (6.4.1) is an entire function, and is therefore $\equiv 0$.

The identities (6.4.2) also follow from the fact that $\{f,w\} = O(z^{-4})$ at $\infty$; see Exercise 4. □

**Corollary 6.4.3.** *Under the assumptions of Theorem 6.4.2, the (extended) function $f_e$ is the quotient of two independent solutions of the equation*

$$\varphi''(z) + q(z)\varphi(z) \;=\; 0, \tag{6.4.4}$$

*where*

$$q(z) \;=\; \frac{1}{4}\sum_{k=1}^{n}\left[\frac{1-\alpha_k^2}{(z-a_k)^2} + \frac{2b_k}{z-a_k}\right]. \tag{6.4.5}$$

An equation of the form (6.4.4), where $q$ has the form (6.4.5), is said to be of *Fuchsian type*, i.e. each point of singularity of $q$, including $\infty$, is a *regular singular point* of the equation. For an equation of second order with meromorphic coefficients $p_j$, this means that near each point $a$, the coefficients of

$$p_2(z)u''(z) + p_1(z)u'(z) + p_0(z)u(z)$$

satisfy

$$\frac{(z-a)\,p_1(z)}{p_2(z)} \;=\; O(1), \qquad \frac{(z-a)^2\,p_0(z)}{p_2(z)} \;=\; O(1).$$

## 6.5 Triangles and hypergeometric functions

From this point on, we drop the subscript and write $f$ for the extension of the inverse of the Riemann map $g$.

In the case of a curvilinear triangle $P$, we may choose the points $a_j$ that map to the vertices to be three arbitrary points of $\mathbb{R}$, or of $\mathbb{R}\cup\{\infty\}$. As is so often the case, one chooses $a_1 = 0$, $a_2 = 1$, $a_3 = \infty$. We need to re-examine the Schwarzian $\{f,z\}$ near the point at $\infty$. If we take $a_3$ to be $\infty$, the behavior (6.4.3) of $f$ near $w = \infty$ is

$$f(w) \;=\; w^{-\alpha_3}\, h\left(\frac{1}{w}\right).$$

Then

$$f'(w) = w^{-\alpha_3-1}\left[-\alpha_3 h\left(\frac{1}{w}\right) - \frac{1}{w}h'\left(\frac{1}{w}\right)\right],$$
$$f''(w) = w^{-\alpha_3-2}(\alpha_3+1)\left[\alpha_3 h\left(\frac{1}{w}\right) + \frac{2}{w}h'\left(\frac{1}{w}\right) + O(w^{-2})\right].$$

It follows from this that as $w \to \infty$,

$$\{f,w\} = \frac{1-\alpha_3^2}{2w^2} + O(w^{-3}).$$

Therefore for each finite $w$,

$$\{f,w\} = \frac{1-\alpha_1^2}{2w^2} + \frac{b_1}{w} + \frac{1-\alpha_2^2}{2(w-1)^2} + \frac{b_2}{w-1}. \tag{6.5.1}$$

In fact the difference between the left and right sides of (6.5.1) is an entire function that vanishes at $\infty$. The coefficients $b_1, b_2$ can be determined by expanding in powers of $1/w$:

$$2b_2 = \alpha_1^2 + \alpha_2^2 - \alpha_3^2 - 1; \qquad b_1 + b_2 = 0.$$

Therefore

$$\{f,w\} = \frac{1-\alpha_1^2}{2w^2} + \frac{1-\alpha_2^2}{2(w-1)^2} + \frac{\alpha_3^2+1-\alpha_1^2-\alpha_2^2}{2w(1-w)}, \tag{6.5.2}$$

and $f$ is the quotient of two independent solutions of the equation

$$\varphi''(w) + \frac{1}{4}\left[\frac{1-\alpha_1^2}{w^2} + \frac{1-\alpha_2^2}{(w-1)^2} + \frac{\alpha_3^2+1-\alpha_1^2-\alpha_2^2}{w(1-w)}\right]\varphi(w) = 0. \tag{6.5.3}$$

This equation has singular points at 0, 1, and $\infty$, each of them regular. There is a standard, much studied, example of such an equation: the Gauss hypergeometric equation

$$z(1-z)u''(z) + [c - z(a+b+1)]u'(z) - ab\,u(z) = 0, \tag{6.5.4}$$

or

$$u''(z) + Au'(z) + Bu(z) = 0, \tag{6.5.5}$$

where

$$A = \frac{c}{z} + \frac{c-(a+b+1)}{1-z}, \quad B = -\frac{ab}{z(1-z)}.$$

We may convert an equation of the form (6.5.5) to the form (6.5.3) by writing $u = v\varphi$ and choosing the function $v$ so as to eliminate the coefficient of $\varphi'$ in the equation for $\varphi$ that results from the equation (6.5.5) for $u$:

$$\varphi'' + \left[\frac{2v'}{v} + A\right]\varphi' + \left[\frac{v''}{v} + A\frac{v'}{v} + B\right]\varphi = 0.$$

Thus we want $v'/v = -A/2$. Then

$$\frac{v''}{v} = \left(\frac{v'}{v}\right)' + \left(\frac{v'}{v}\right)^2 = -\frac{A'}{2} + \frac{A^2}{4}, \quad \frac{v'}{v}A = -\frac{A^2}{2},$$

which leads to the equation

$$\varphi'' + \left[-\frac{A'}{2} - \frac{A^2}{4} + B\right]\varphi = 0. \tag{6.5.6}$$

With $A$ and $B$ in (6.5.5), the coefficient of $\varphi$ is

$$\frac{2c-c^2}{4z^2} + \frac{2d-d^2}{4(z-1)^2} + \frac{2cd-4ab}{4z(1-z)}, \quad d = a+b+1-c.$$

To get (6.5.3) we want

$$(1-c)^2 = \alpha_1^2, \quad (1-d)^2 = \alpha_2^2, \quad 2cd-4ab = \alpha_3^2+1-\alpha_1^2-\alpha_2^2. \tag{6.5.7}$$

Taking $c = 1-\alpha_1$, $d = 1-\alpha_2$, we find that $a+b = 1-(\alpha_1+\alpha_2)$. Then

$$\begin{aligned}\alpha_3^2+1-\alpha_1^2-\alpha_2^2 &= \alpha_3^2+1-(1-c)^2-(1-d)^2\\ &= \alpha_3^2+1-(1-c)^2-(a+b-c)^2\\ &= \alpha_3^2-2c^2-2c+2c(a+b)-(a+b)^2\\ &= \alpha_3^2-2c(c-a-b-1)-(a-b)^2-4ab\\ &= \alpha_3^2+(2cd-4ab)-(a-b)^2.\end{aligned}$$

Thus we may satisfy all the equations of (6.5.7) by taking

$$a+b = 1-(\alpha_1+a_2), \quad a-b = \alpha_3, \quad c = 1-\alpha_1.$$

We do not need an explicit determination of the factor $v$ (which is easily calculated), since $v$ drops out when we take the quotient $\varphi_0/\varphi_1 = vu_0/vu_1$, where the $u_j$ are solutions of the hypergeometric equation (6.5.5).

The point of all this is that the solutions of (6.5.4) are very well-understood. For example, (6.5.4) may be written in the form

$$[D(D+c-1)-z(D+a)(D+b)]u(z) = 0, \qquad D = z\frac{d}{dz}. \tag{6.5.8}$$

Since $D[z^\nu] = \nu z^\nu$, the series expansion of a solution with $u(0) = 1$ is easily calculated:

$$u_0(z) = F(a,b;c,z) \equiv \sum_{n=0}^{\infty} \frac{(a)_n(b)_n}{(c)_n n!} z^n, \quad c \neq 0,-1,-2,\ldots, \tag{6.5.9}$$

where $(a)_n = a(a+1)\cdots(a+n-1)$, etc. There are also representations of this function as an integral; for such a representation due to Euler see Exercise 15.

The map $z \to 1-z$ takes a solution of (6.5.5) to a solution of the equation with different parameters. In particular, a second solution of (6.5.4), regular at $z=1$, is

$$u_1(z) \;=\; F(a,b;a+b+1-c;1-z) \;=\; F(a,b;d;1-z). \tag{6.5.10}$$

By Theorem 6.3.1, the ratio $u_0/u_1$ is related to the map $f$ by a linear fractional transformation. Therefore it is the mapping function for a corresponding curvilinear triangle. The original $F$ can be recovered by choosing the linear fractional transformation $g$ so that $g \circ (u_0/u_1)$ maps $\{0,1,\infty\}$ to $\{f(0), f(1), f(\infty)\}$, respectively.

**Remark.** For ordinary triangles (those with straight sides), the continuation of the mapping function is itself single-valued in certain restricted cases (see Section 5.6). The same is true, with much milder restrictions, for curvilinear triangles; see Exercise 10.

## 6.6 Regular polygons and hypergeometric functions

Up to an affine transformation, a *regular* curvilinear polygon is one whose vertices are equally spaced on the unit circle, say at the points $z_k = \omega^k = \exp(2\pi ik/n)$, $k = 0,1,\dots,n-1$, and all of whose angles are equal, say $\pi\alpha$. A map $f : \mathbb{D} \to P$ can be chosen in such a way that $f$ commutes with rotation through an angle $\omega = 2\pi/n$:

$$f(\omega z) \;=\; \omega f(z), \tag{6.6.1}$$

and so that $f(\omega^k) = \omega^k$, see Problem 3 of Chapter 5. We may take advantage of this symmetry to convert the corresponding equation of Fuchsian type to the hypergeometric equation, and express $f(z)$ as the quotient of two hypergeometric functions of $z^n$.

The original equation has the form

$$\varphi''(z) + q(z)\varphi(z) \;=\; 0, \qquad q(z) \;=\; \frac{\{f,z\}}{2}, \tag{6.6.2}$$

where

$$\{f,z\} \;=\; \frac{1}{2}\sum_{k=0}^{n-1}\left[\frac{1-\alpha^2}{(z-\omega^k)^2} + \frac{2b_k}{z-\omega^k}\right], \qquad \alpha = \frac{\omega}{\pi} = \frac{2}{n}. \tag{6.6.3}$$

Let $f_1(z) = f(\omega z)$. Then it is easily seen that the Schwarzian satisfies

$$\{f_1,z\} \;=\; \omega^2\{f,\omega z\}. \tag{6.6.4}$$

On the other hand, (6.6.1) and the invariance properties of the Schwarzian imply that

$$\{f_1,z\} \;=\; \{\omega f,z\} \;=\; \{f,z\}. \tag{6.6.5}$$

Combining (6.6.4) and (6.6.5) we see that $z^2\{f,z\}$ is invariant under rotation by an angle $\alpha = 2\pi/n$. Putting the sum in (6.6.3) over a common denominator, and using the rotation invariance, we find that

$$z^2\{f,z\} = \frac{Q(z^n)}{2(z^n-1)^2},$$

where $Q$ is a polynomial of degree less than 2, with no constant term: $Q(w) = aw$. Near $z = 1$ we have, for some constant $a$,

$$z^2\{f,z\} = \frac{a\,z^n}{2(z^n-1)^2} = \frac{(1-\alpha^2)\,z^2}{2(z-1)^2} + O\left(\frac{1}{z-1}\right).$$

Multiplying both sides by $(z-1)^2$ and taking the limit as $z \to 1$, we find that $a = n^2(1-\alpha^2)$. Thus

$$z^2\{f,z\} = \frac{n^2(1-\alpha^2)\,z^n}{2(z^n-1)^2}. \tag{6.6.6}$$

Let us write a solution $\varphi$ of (6.6.2) as $\varphi(z) = \psi(z^n)$. Then equation (6.6.2), multiplied by $z^2$ is

$$\begin{aligned} 0 &= z^2[\psi(z^n)]'' + \frac{z^2\{f,z\}}{2}\,\psi(z^n) \\ &= n^2z^{2n}\psi''(z^n) + n(n-1)z^n\psi'(z^n) + z^2\frac{\{f,z\}}{2}\,\psi(z^n). \end{aligned}$$

Taking into account (6.6.6), we may change variables and write this as

$$\psi''(z) + \frac{\beta}{z}\,\psi'(z) + \frac{1-\alpha^2}{4z(1-z)^2}\,\psi(z) = 0, \qquad \beta = 1 - \frac{1}{n}. \tag{6.6.7}$$

As in the previous section, we seek to transform this into a hypergeometric equation (6.5.5) by writing $u = v\psi$. As before the resulting equation for $\psi$ is

$$\psi'' + \left(\frac{2v'}{v} + A\right)\psi' + \left(\frac{v''}{v} + \frac{v'}{v}A + B\right)\psi = 0. \tag{6.6.8}$$

In this case, we want to choose $v$ so that

$$\frac{2v'}{v} + A = \frac{\beta}{z}.$$

Then

$$\begin{aligned} \frac{v''}{v} &= \left(\frac{v'}{v}\right)' + \left(\frac{v'}{v}\right)^2 = -\frac{A'}{2} - \frac{\beta}{2z^2} + \frac{1}{4}\left(A - \frac{\beta}{z}\right)^2; \\ \frac{v'}{v}A &= \left(\frac{\beta}{2z} - \frac{A}{2}\right)A. \end{aligned}$$

As in the calculation in the previous section, the lowest order coefficient in the equation (6.6.8) is

$$-\frac{A'}{2}-\frac{\beta}{2z^2}+\frac{1}{4}\left(A-\frac{\beta}{z}\right)^2-\frac{A^2}{2}+\frac{\beta A}{2z}+B \tag{6.6.9}$$

$$=\frac{(\beta-1)^2-(c-1)^2}{4z^2}+\frac{1-(d-1)^2}{4(1-z)^2}+\frac{2cd-4ab}{4z(1-z)}. \tag{6.6.10}$$

The resulting equation is

$$\frac{(\beta-1)^2-(c-1)^2}{4z^2}+\frac{1-(d-1)^2}{4(1-z)^2}+\frac{2cd-4ab}{4z(1-z)} = \frac{1-\alpha^2}{4z(1-z)^2}. \tag{6.6.11}$$

A look at the behavior as $z \to 0$ and as $z \to 1$ shows that $(c-1)^2 = (\beta-1)^2 = 1/n^2$, so (6.6.11) reduces to

$$\frac{1-(d-1)^2}{1-z}+\frac{2cd-4ab}{z} = \frac{1-\alpha^2}{z(1-z)} = \frac{1-\alpha^2}{1-z}+\frac{1-\alpha^2}{z}.$$

Therefore

$$1-(d-1)^2 = 2cd-4ab = 1-\alpha^2.$$

Since $d-1 = a+b-c$, we may take $a+b = c-\alpha$. Then $d = 1-\alpha$ and

$$(a-b)^2 = (a+b)^2-4ab = (c-\alpha)^2-2cd+1-\alpha^2 = (c-1)^2 = \frac{1}{n^2}.$$

The associated hypergeometric equation can be taken to have indices

$$c = 1-\frac{1}{n}, \quad a = \frac{1-\alpha}{2}, \quad b = \frac{1-\alpha}{2}-\frac{1}{n}. \tag{6.6.12}$$

Equation (6.5.4) with these indices has solutions

$$u_0(z) = C z^{1/n} F(a+\tfrac{1}{n}, b+\tfrac{1}{n}; 1+\tfrac{1}{n}; z), \tag{6.6.13}$$

$$u_1(z) = F(a, b; 1-\alpha; 1-z), \tag{6.6.14}$$

where the constant

$$C = \frac{\Gamma(\frac{1}{2}[1+\alpha])\,\Gamma(\frac{1}{2}[1+\alpha+\frac{2}{n}])}{\Gamma(1+\frac{1}{n})\,\Gamma(\alpha)}$$

is chosen so that $u_0(1) = 1$; see Exercise 16.

**Theorem 6.6.1.** *Up to a rotation, the mapping function $f : \mathbb{D} \to P$, where $P$ is a regular curvilinear polygon with $n$ vertices and angle $\pi\alpha$ is*

$$f(z) = \frac{u_0(z^n)}{u_1(z^n)}, \tag{6.6.15}$$

*where the $u_j$ are the hypergeometric functions (6.6.13), (6.6.14), with the indices (6.6.12).*

*Proof:* The constant $C$ was chosen so that $f$ and the right side of (6.6.15) agree at $z = 1$. Rotation invariance of $f$ implies that they agree at each of the points $\omega^k$, $0 \leq k \leq n-1$. Since $n \geq 3$, they are identical. $\square$

## Exercises

1. Use the function $\langle z_0, z_1, z_2 \rangle$ of Exercise 1 of Chapter 2 to derive a measure of the deviation of a holomorphic function $f : \mathbb{C} \to \mathbb{C}$ from an affine map, in analogy with the derivation of the Schwarzian derivative.
2. Complete the details in the proof of Proposition 6.2.1.
3. Suppose that $f$ is holomorphic at $\infty$, with non-vanishing derivative:
$$f(z) \ = \ a_0 + az^{-1} + bz^{-2} + O(z^{-3}), \quad a \neq 0.$$
Show that $\{f, z\} = O(z^{-4})$.
4. Confirm that the identities (6.4.2) follow from the fact that $\{f, z\} = O(z^{-4})$ as $z \to \infty$.
5. The case of a curvilinear polygon with $n = 2$ is a crescent. Note that the two angles $\alpha$ must be equal. Derive the form of the associated second-order equation when the points $a_1$, $a_2$ are taken to be 0, $\infty$.
6. The general second-order Fuchsian equation with singular points only at 0, $\infty$ can be shown to have the form $(D-a)(D-b)u \ = \ 0$, where again $Du(z) = zu'(z)$. Show that if $a \neq b$, there are two independent solutions that are powers of $z$.
7. Show that an equivalent form for the equation in Exercise 6 is
$$u''(z) + \frac{1-a-b}{z} u'(z) + \frac{ab}{z^2} u(z) \ = \ 0.$$
Deduce that conditions on $a, b$ that put this equation into the form of the equation in Exercise 5 are $a+b=1$, $4ab = 1-\alpha^2$. Solving these equations gives two independent solutions of the equation in Exercise 5, and thus a representation of a map to the image of the crescent under some linear fractional transformation.
8. A more direct way to find a map to the crescent with angles $\alpha$: given such a map, by applying a linear fractional transformation we may assume that one vertex is at 0 and the other at $\infty$. Then the two sides must be straight lines through the origin that meet at an angle $\alpha$. Compare this with the result in Exercise 7.
9. Several domains that appear in Chapter 17 are bounded below by a circular arc and on the sides by vertical lines. Such a domain can be considered as a curvilinear triangle having a vertex at infinity with angle zero, and two equal angles $\alpha$ at the finite vertices. What are the corresponding hypergeometric equations?

10. Show that the continuation of the mapping function of a curvilinear triangle is single-valued if and only if the angle at each vertex (separately) is $\pi/m$, where $m$ is a positive integer.
11. Use the rotation invariance of $z^2\{f,z\}$ and the determination of the constant $a$ in (6.6.6) to find the constants $b_k$ in (6.6.3).
12. Discuss the limiting case of a regular curvilinear polygon with $n$ vertices when $\alpha \to 1$.
13. Discuss the case of a regular curvilinear polygon when $\alpha \to 0$.
14. What are the angles for the case of a regular polygon with $n$ vertices and straight sides. What is the corresponding hypergeometric equation?
15. For results about the gamma and beta functions used in this exercise, see Chapter 10. Assume here that $a, b, c$ all have positive real part.

    (a) Use the identities $(b)_n = \Gamma(b+n)/\Gamma(b)$ and $\Gamma(a)\Gamma(b)/\Gamma(a+b) = \mathrm{B}(a,b)$ to write the coefficient of $z^n$ in (6.5.9) as

$$\frac{\Gamma(c)\Gamma(b-c)}{\Gamma(b)}\,(a)_n\,\mathrm{B}(b+n, c-b).$$

    (b) Use the integral representation of the beta function,

$$\mathrm{B}(\alpha,\beta) \;=\; \int_0^1 s^{\alpha-1}(1-s)^{\beta-1}\,ds,$$

    to convert the sum (6.5.9) to *Euler's integral formula*

$$\begin{aligned} F(a,b;c;z) &= \frac{1}{\mathrm{B}(b,c-b)}\int_0^1 s^{b-1}(1-s)^{c-b-1}\sum_{n=0}^{\infty}\frac{(a)_n}{n!}\,(sz)^n\,ds \\ &= \frac{1}{\mathrm{B}(b,c-b)}\int_0^1 s^{b-1}(1-s)^{c-b-1}(1-sz)^{-a}\,ds. \end{aligned} \tag{6.6.16}$$

16. (a) Under the assumptions of Exercise 15, if also $\mathrm{Re}\,(c-a-b) > 0$, prove *Gauss's evaluation*

$$F(a,b;c;1) \;=\; \frac{\Gamma(c)\,\Gamma(c-a-b)}{\Gamma(c-a)\,\Gamma(c-b)}.$$

    (b) Formally part (a) shows that $F(a,b;c;1) = \infty$ if $a+b=c$. Show that this follows from (6.6.16).
17. Use the form (6.5.8) and the fact that $z^{-\alpha}D[z^{\alpha}f(z)] \;=\; (D+\alpha)f(z)$, to show that the function

$$z^{1-c}F(a+1-c, b+1-c; 2-c; z)$$

    is a solution of equation (6.5.5).
18. The elliptic modular function $\lambda$ of Chapter 17 is the conformal map from $\mathbb{C}_+$ to a curvilinear triangle with vertices $0, 1, \infty$ and angles $\alpha = 0$; $\lambda(\infty) = 0$, $\lambda(0) = 1$, $\lambda(1) \;=\; \infty$. Use the results of Section 6.5 and Exercises 16 and 17 to show that $\lambda$ is a constant multiple of the quotient

$$\frac{(z)^{1/2}F(1,1;3/2;z)}{(1-z)^{1/2}F(1,1;3/2;1-z)}.$$

## Remarks and further reading

For more detail on the Schwarzian and its history, see Hille [64], [65]. For more on the geometric interpretation, see Ovsienko and Tabachnikov [113], [114]. For a fuller discussion of the mapping of a curvilinear triangle, see Ablowitz and Fokas [2].

The Schwarzian has a connection to univalent functions. If a holomorphic function defined in $\mathbb{C}_+$ is univalent, then $y^2|\{f,z\}| \leq 3/2$. Conversely, if $y^2|\{f,z\}| \leq 1/2$, then $f$ is univalent. See Nehari [103].

The Schwarzian also plays a role in the approach of Bers to Teichmüller spaces; see Bers [22] and Hubbard [68], Chapter 6.

# Chapter 7
# Riemann surfaces and algebraic curves




A Riemann surface can be thought as the domain of definition of a holomorphic function $f$ that has been continued analytically as far as such continuations can be carried out. In general this is not a domain in the previous sense, i.e. a subset of the plane. Rather it is a complex manifold of one (complex) dimension that projects locally into $\mathbb{C}$.

The functions most often considered in connection with their Riemann surfaces are those associated with *algebraic curves*. Such a function $f$ is defined implicitly by an equation

$$P(z, f(z)) = 0, \tag{7.0.1}$$

where $P(z,w)$ is an irreducible polynomial in two variables $z, w$. For example, the polynomial $P(z,w) = w^2 - z$ leads to the function $f(z) = \sqrt{z}$.

The equation $P(z,w) = 0$ defines a curve in $\mathbb{C}^2$, or in $\mathbb{S}^2$, the subset $C = \{(z,w)\}$ of pairs $(z,w)$ that satisfy the equation. As we shall see, $C$ can be identified with the Riemann surface of $f$.

After examining the process of analytic continuation, and defining Riemann surfaces in general, we show that the Riemann surface of an analytic function $f$ is compact if and only if $f$ satisfies an equation (7.0.1), where $P$ is an irreducible polynomial. This requires some study of both the algebraic properties and the analytic properties of a polynomial in two complex variables.

There is some overlap between this chapter and Chapters 15 and 16 on elliptic functions, but the presentations are independent.

## 7.1 Analytic continuation

We start with the notion of a *function element*. This is a pair $[f, D]$ that consists of an open non-empty disk

$$D_r(z_0) = \{z : |z - z_0| < r\}, \quad 0 < r < \infty,$$



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_7

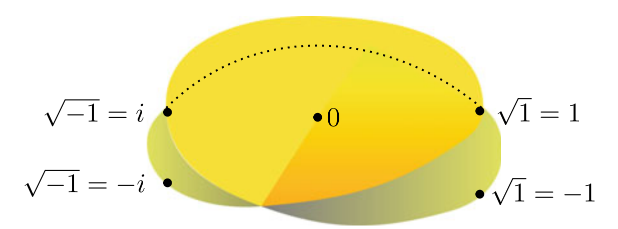



**Fig. 7.1** A portion of the continuation of $\sqrt{z}$

and a holomorphic function $f$ defined on $D_r(z_0)$, e.g. by a convergent series

$$f(z) \;=\; \sum_{n=0}^{\infty} a_n(z-z_0)^n, \quad |z-z_0| < r.$$

A *chain* of function elements is a collection of function elements $\{[f_j, D_j]\}_{j=0}^{n}$ such that successive disks overlap and the corresponding functions agree on the overlap:

$$D_j \cap D_{j+1} \;\neq\; \emptyset; \quad f_j \;=\; f_{j+1} \;\text{ on }\; D_j \cap D_{j+1}, \;\; j = 0, 1, 2, \ldots, n-1.$$

As an example, consider the square-root function $z^{1/2}$. Starting with a point $x > 0$ and a branch of $z^{1/2}$ holomorphic for $|z-x| < x$, consider a chain of overlapping disks with radii $x$ and centers $z_k$ on the circle $\{z : |z| = x\}$, such that the arguments of the $z_k$ increase. Suppose that $\arg z_n = 2\pi$. Then the domains of the first and last function elements are the same, but the values may differ by a factor $-1$, see Figure 7.1.

This example illustrates an important question: Suppose that a chain of function elements $\{[f_j, D_j]\}_{j=0}^{n}$ returns to its starting point; does $f_n = f_0$? A useful notion in this connection is *continuation along a curve*, as introduced in Section 1.7. We say that a function element $[f_1, D_1]$ can be continued along a curve $\gamma$ if there is a chain such that the union of the $D_j$ contains $\gamma$. (We use the same notation $\gamma$ for the map $\gamma : I \to \mathbb{C}$, where $I$ is some real interval, and for its image $\gamma(I)$ in $\mathbb{C}$.) We may always take such a curve to be a polygonal line connecting the centers of the disks.

The monodromy theorem, Theorem 1.7.2, is the basis for the development of these ideas. For convenience we repeat it here, in slightly different phrasing. We begin with a consequence of the uniqueness result, Proposition 1.7.1.

**Proposition 7.1.1.** *Any two continuations of a function element along a curve are identical in a neighborhood of the curve.*

Recall that a domain $\Omega \subset \mathbb{C}$ is said to be *simply connected* if each closed curve in $\Omega$ can be deformed continuously to a point (a constant curve).

**Theorem 7.1.2.** (Monodromy theorem) *Suppose that the domain $\Omega$ is simply connected. Suppose that $[f_0, D_0]$ is a function element that can be continued along each curve in $\Omega$. Then $f_0$ has a unique holomorphic extension to all of $\Omega$.*

A case where the theorem does not apply is a punctured disk:

$$\Omega \;=\; D_r(z_0)\setminus\{z_0\} \;=\; \{z : 0 < |z-z_0| < r\}.$$

Suppose that $f$ can be continued along each curve in $\Omega$. In particular we consider continuation along circles centered at $z_0$, in the positive direction. Suppose some such continuation returns to the original function element in $k$ circuits. Proposition 7.1.1 implies that the same is true for each such continuation. The logarithm at $z_0 = 0$ is an example for which there is no return.

If the return number $k = 1$, then $z_0$ is an isolated singularity for $f$. If it is removable, then $[f, D]$ is a function element. If it is a pole, we may associate to it the Laurent expansion

$$\sum_{n=N}^{\infty} a_n (z-z_0)^n.$$

Suppose now that the first return is after $k$ circuits, $k > 1$. Suppose that the original function element is defined in a disk $D_{r/2}(z_0 + a)$, where $|a| = r/2$. Define

$$g(w) \;=\; f(z_0 + w^k), \qquad 0 < |w| < a^{1/k}.$$

This relation can be continued around a circle centered at $w = 0$. As $w$ makes one circuit around 0, $z_0 + w^k$ makes $k$ circuits around $z_0$. Therefore $g$ returns to its original value. Thus $g$ is single-valued in a punctured disk centered at 0. If $z_0$ is either a pole or a removable singularity for $f$, then there is an expansion

$$f(z) \;=\; f(z_0 + (z-z_0)) \;=\; g\left((z-z_0)^{1/k}\right) \;=\; \sum_{n=N}^{\infty} a_n (z-z_0)^{n/k}. \tag{7.1.1}$$

This expansion is valid for the original function element, for a particular branch of the $k$-th root, and it remains valid as $f$ is continued along circles centered at $z_0$. The expansion (7.1.1) is called the *Puiseux expansion* of $f$ at $z_0$. There are $k$ distinct branches of $f$ over each nearby point $z \neq z_0$.

Similar considerations apply at $z = \infty$, if some function element can be continued throughout a neighborhood $\{z : |z| > R\}$. This leads to the possibility of a Taylor, Laurent, or Puiseux expansion

$$f(z) \;=\; \sum_{n=N}^{\infty} a_n z^{-n/k}$$

for some integer $k \geq 1$.

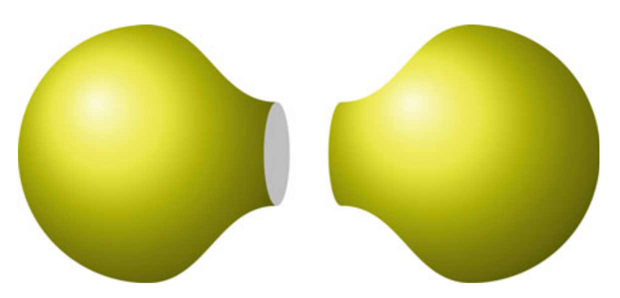


**Fig. 7.2** Joining across a slit

## 7.2 The Riemann surface of a function

Figure 7.1 above can be considered an illustration of the portion of the Riemann surface of the $\sqrt{z}$ that lies above the unit disk $\mathbb{D}$. For an illustration of a portion of the Riemann surface of the logarithm; see Figure 1.2.

Visualizing an entire surface, with its topology, requires a little more thought. The basic idea can be illustrated by returning to the square-root function. Except for $z = 0$, the equation $w^2 = z$ has two solutions. The domain obtained by cutting the plane along the negative real axis,

$$\Omega \;=\; \mathbb{C}\setminus(-\infty, 0],$$

is simply connected. We take two sheets—copies of the slit plane—and join them across the slit in such a way that proceeding in the positive direction around the origin takes us from either sheet to the other sheets. The resulting continuation of $f(z) = \sqrt{z}$ is single-valued and holomorphic on the resulting set, except at the points $z = 0$ and $z = \infty$, which belong to both sheets. If we add in these two points, the result is topologically a sphere.

This procedure is most easily pictured by starting from two copies of the Riemann sphere, opening each along the slit, and gluing together at the resulting curves; see Figure 7.2. Thus the "complete analytic function" $f(z) = z^{1/2}$ is defined on this sphere minus the two points that correspond to $z = 0$ and $z = \infty$, while the Riemann surface $R_f$ is the complete sphere.

This sphere can be given the structure of a *complex manifold* in such a way that $f$ is a meromorphic function on $R_f$. In fact each function element $[f, D_r(z_0)]$ gives us an open set with a coordinate function $z - z_0$; on overlapping disks $D$ the change from one coordinate to the other on the overlap is holomorphic (in fact linear). At $z_0 = 0$ we take $z^{1/2}$ as the coordinate. It is single-valued—one circuit around the origin takes us from one sheet to the other. At $z = \infty$ we take $(1/z)^{1/2}$ as a coordinate. Again, coordinate changes on overlaps are holomorphic. Thus $R_f$ is a complex manifold. Defined in the obvious way, $f$ is meromorphic, with a simple zero at 0 and a simple pole at $\infty$.

There are a number of different ways to set up the general version of this construction, and terminology varies. The following construction and terminology is, in our view, the clearest and most convenient.

A *regular point* is a pair $|z_0, f|$ consisting of a point $z_0$ in the Riemann sphere $\mathbb{S}$ and, if $z_0 \neq \infty$, a power series

$$f(z) = \sum_{n=0}^{\infty} a_n (z-z_0)^n \tag{7.2.1}$$

that converges in some disk $D$ of positive radius centered at $z_0$. It is assumed that not all coefficients are zero. If $z_0 = \infty$ the series has the form

$$\sum_{n=0}^{\infty} a_n z^{-n}$$

and the disk has the form $\{z : |z| > r\}$. The number $z_0$ is said to be the *base point* of the regular point.

A *polar point* is a pair $|z_0, f|$ consisting of a point $z_0$ and a Laurent series

$$f(z) = \sum_{n=N}^{\infty} a_n (z-z_0)^n, \qquad N < 0, \; a_N \neq 0, \tag{7.2.2}$$

that converges in some punctured disk $D \setminus \{z_0\}$ centered at $z_0$. For $z_0 = \infty$ the expansion is in powers $(1/z)^n$.

An *algebraic branch point* is a pair $|z_0, f|$ consisting of a point $z_0$ and a Puiseux series

$$f(z) = \sum_{n=N}^{\infty} a_n (z-z_0)^{n/k}, \qquad a_N \neq 0, \; k > 1, \tag{7.2.3}$$

such that the associated series $\sum_{n=N}^{\infty} a_n w^n$ converges for $w$ in some punctured disk centered at $w = 0$. For $z_0 = \infty$ the terms are $a_n (1/z)^{n/k}$. We assume that the expansion (7.2.3) is not such that only powers of $(z - z_0)^{1/l}$ have non-zero coefficients, for some $l < k$. Then the first return of $f$ to the same values is after $k$ circuits. The *multiplicity* of the branch point is the integer $k$. Consistent with this definition, we take the multiplicity of a regular point or a polar point to be 1.

A regular point $|z_0, f|$ defines a function element $[f, D]$ in the previous sense, with $D$ a disk centered at $z_0$ and $f$ the function defined by the associated series (7.2.1). The disk $D$ can be taken to have radius equal to the radius of convergence of the series. Two regular points are said to be *related* if there is a chain of overlapping function elements that links one to the other. This is clearly an equivalence relation. The only modification needed in the case of polar points and branch points is to broaden our definition of "function element" to include pairs consisting of punctured disks and the functions defined in them by Laurent expansions (7.2.2) or Puiseux expansions (7.2.3). We do so, and define a *Riemann surface* $R_f$ to consist of all the points—regular points, polar points, and algebraic branch points—that are related to some one regular point $|z_0, f|$

As in the case of the square-root function, analytic continuation can lead to several points of $R_f$ that have the same base point $z_0$. In the case of regular or polar points, each such point is said to be a *branch* at $z_0$. In the case of a branch point of $R_f$ with base point $z_0$ and expansion (7.2.3), each $z$ sufficiently close to $z_0$ is the base point for $k$ regular points, each of them close to the branch point in $R_f$.

A Riemann surface $R_f$ can be made a complex manifold exactly as in the case of the square-root function. Each disk $D_{z_0}$ comes with a coordinate: $z - z_0$ or $1/z$ in the case of a regular or polar point, and $(z-z_0)^{1/k}$ or $z^{-1/k}$ in the case of an algebraic branch point. Two such disks are disjoint if the associated functions $f$ disagree on their entire overlap, and are linked if their intersection is non-empty and the associated functions agree on the intersection. The function $f$ is defined, single-valued, and meromorphic on $R_f$.

The map $|z, f| \to (z, f(z)) \in \mathbb{S}^2$ identifies the Riemann surface $R_f$ with the curve defined by the equation $w = f(z)$.

## 7.3 Compact Riemann surfaces

The following is the first half of a two-way connection between compact Riemann surfaces and algebraic functions.

**Theorem 7.3.1.** *Suppose that the Riemann surface $R_f$ is compact. Then $\{(z, f(z))\}$ is an algebraic curve: there is a polynomial $P(z,w)$ such that $P(z, f(z)) \equiv 0$ on the surface $R_f$.*

*Proof:* The poles and branch points of $R_f$ are isolated points, so they are finite in number. Let $\{z_j\}_{j=1}^{d}$ be the collection of base points of the poles and branch points. Choose another point $z_0 \in \mathbb{C}$ and let $\gamma_j$ be a family of non-intersecting simple curves with $\gamma_j$ running from $z_0$ to $z_j$. Let $\Omega$ be the complement in $\mathbb{S}$ of the union of the curves $\gamma_j$. Then $\Omega$ is simply connected. Suppose $f$ is holomorphic in some neighborhood of a point $w_0 \in \Omega$. Let $w$ be any other point of $\Omega$, and let $\gamma$ be a curve from $w_0$ to $w$. The set of points along $\gamma$ to which $f$ can be analytically continued from $w_0$ is clearly open relative to the curve. On the other hand, because of compactness of the Riemann surface, this set is also closed. In fact if $w_n$ is a sequence of such points, some subsequence of the sequence of branches $\{|w_n, f|\}$ converges. Therefore $f$ extends to each point of $\Omega$. By the monodromy theorem, it is single-valued.

We have shown that, to each point of the Riemann surface whose base point $w_0$ is in $\Omega$, there corresponds a copy of $\Omega$ and a single-valued determination of $f$ on that copy. There are only finitely many such copies. In fact, choose a $w_0 \in \Omega$. The points of the surface with base point $w_0$ but different determinations of $f$ are isolated in $R_f$, so there are only finitely many points with base point $w_0$. The number $m$ of such points is constant along each curve in $\Omega$, and therefore is constant throughout $\Omega$. Thus $R_f$ consists of $m$ sheets—$m$ copies of $\Omega$—joined in some way across the curves $\gamma_j$.

Number the sheets, and for each $z \in \Omega$ let $\{|z, f_j|\}_{j=1}^m$ be the corresponding points of the Riemann surface. Consider the function

$$Q(z,w) \;=\; \prod_{j=1}^{m}[w - f_j(z)] \;=\; \sum_{k=1}^{m}(-1)^{m-k}S_{m-k}(z)\,w^k,$$

where the $S_k$ are the elementary symmetric polynomials in the $f_j$:

$$S_0 \;=\; 1, \quad S_1 \;=\; \sum_{j=1}^{m} f_j, \quad S_2 \;=\; \sum_{j\neq k} f_j f_k, \quad \ldots \quad, \quad S_m \;=\; \prod_{j=1}^{m} f_j.$$

Because of their symmetry, the functions $S_j$ are single-valued at points $z \in \Omega$. Moreover there was nothing special about the particular point $z_0$ or the particular choice of curves $\{\gamma_j\}$, so the $S_j$ are single-valued holomorphic functions in the complement of the points $\{z_j\}_{j=1}^d$. Each of the points $z_j$, $j = 1, 2, \ldots, d$, is an isolated singularity at which each $S_k$ grows at most like a negative power of $|z - z_j|$. Therefore each $S_j$ is a rational function of $z$.

Let $q(z)$ be the least common multiple of the denominators of the rational functions $S_0, S_1, \ldots, S_{m-1}$. The functions

$$q_k(z) \;=\; (-1)^{m-k}q(z)S_{m-k}(z)$$

are polynomials. Thus

$$P(z,w) \;=\; q(z)\prod_{j=1}^{m}[w - f_j(z)] \;=\; \sum_{k=0}^{m} q_k(z)w^k$$

is a polynomial in $z$ and $w$. By construction, $P(z,w)$ vanishes whenever $w = f(z)$ for some point $|z, f|$ in the Riemann surface of $f$. □

As we shall see below, the polynomial $P$ in Theorem 7.3.1 is *irreducible*: it is not the product of two non-constant polynomials in $z, w$. The goal of the next two sections is to prove the converse: if $P(z,w)$ is an irreducible polynomial of positive degree in $w$, then the Riemann surface of the function $f$ defined by

$$P(z, f(z)) \;=\; 0$$

is compact.

## 7.4 Algebraic curves: some algebra

In principle, if $P(z,w)$ has real coefficients, we could consider the algebraic curve $P(x,y) = 0$ in $\mathbb{R}^2$. The example $x^2 + y^2 + 1 = 0$ shows why algebraic curves are more often studied as subsets of $\mathbb{C}^2$.

Here again we restrict the term "algebraic curve" to sets

$$C = \{(z,w) : P(z,w) = 0\}, \tag{7.4.1}$$

where $P$ is irreducible. The polynomial $P$ can be written in the form

$$P(z,w) = q_n(z)\,w^n + q_{n-1}(z)\,w^{n-1} + \cdots + q_0(z), \quad n > 0. \tag{7.4.2}$$

Here each $q_k$ is a polynomial, $q_n$ is not identically zero, and at least one of the $q_k$ is not constant. These conditions guarantee that the curve $C$ is not empty and that the solutions $(z,w)$ depend on $z$. Much more can be said, but to say it we need to start with some purely algebraic considerations.

The ring of complex polynomials in one variable is denoted by $\mathbb{C}[z]$. If $p$ and $q$ are two such polynomials,

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_0,$$
$$q(z) = b_m z^m + b_{m-1} z^{n-1} + \cdots + b_0, \quad a_n b_m \neq 0, \ n \geq m,$$

then there are polynomials $c$ and $r$ such that $\deg c = n - m$, and

$$p(z) = c(z)\,q(z) + r(z), \quad \deg r < \deg q. \tag{7.4.3}$$

In fact (7.4.3) is a system of $n-m+1$ linear equations for the $n-m+1$ coefficients of $c$ that can be solved sequentially. At the first step we choose the leading coefficient $c_{n-m}$ equal to the quotient $a_n/b_m$ of the leading coefficients of $p$ and $q$. At the next step we want

$$a_{n-1} = c_{n-m-1}b_m + c_{n-m}b_{m-1} = c_{n-m-1}b_m + \frac{a_n}{b_m}b_{m-1},$$

which determines $c_{n-m-1}$, and so on. The coefficients of $c$ and of $r$ in (7.4.3) are rational functions of the coefficients of $p$ and $q$.

**Lemma 7.4.1.** *Suppose that $p$ and $q$ are complex polynomials with no common factor. Then there are polynomials $r$ and $s$ such that*

$$r(z)\,p(z) + s(z)\,q(z) \equiv 1. \tag{7.4.4}$$

*Proof:* We may assume that $\deg p \geq \deg q$. Set $p_1 = p$, $p_2 = q$, and carry out the division algorithm (7.4.3) to obtain terms $p_k$ of strictly decreasing degree:

$$p_1 = c_2 p_2 + p_3;$$
$$p_2 = c_3 p_3 + p_4;$$
$$\cdots$$
$$p_{m-3} = c_{m-2} p_{m-2} + p_{m-1};$$
$$p_{m-2} = c_{m-1} p_{m-1} + p_m,$$

where $p_m$ is constant. If $p_m = 0$, then $p_{m-1}$, which has positive degree, divides $p_{m-2}$. The preceding equation shows, then, that $p_{m-1}$ divides $p_{m-3}$. Continuing up

the chain of equations, $p_{m-1}$ divides both $p_2 = q$ and $p_1 = p$, a contradiction. Thus $p_m$ is a constant $a \neq 0$. Starting with

$$\begin{aligned} a &= p_{m-2} - c_{m-1}p_{m-1} \;=\; p_{m-2} - c_{m-1}[p_{m-3} - c_{m-2}p_{m-2}] \\ &= [1 - c_{m-2}]p_{m-2} - c_{m-1}p_{m-3}, \end{aligned}$$

we proceed up the previous chain of equations to reach $p_2 = q$ and $p_1 = p$. Dividing the resulting equation by $a$ gives us (7.4.4). □

We have already spoken of irreducibility for polynomials in two variables. The same concept applies in one variable: $p \in \mathbb{C}[z]$ is *irreducible* if it is not the product of two non-constant elements of $\mathbb{C}[z]$.

**Lemma 7.4.2.** *If an irreducible complex polynomial that belongs to $\mathbb{C}[z]$ divides the product of two polynomials that belong to $\mathbb{C}[z]$, then it divides one (or both) of the factors.*

*Proof:* Suppose that $p$ is irreducible and does not divide either of $q_1$ or $q_2$. Then it has no common factors with either, so there are polynomials $r_j$, $s_j$ such that $r_j p + s_j q_j = 1$. Then

$$s_1 s_2 q_1 q_2 \;=\; (1 - r_1 p)(1 - r_2 p) \;=\; r_1 r_2 p^2 - (r_1 + r_2)p + 1.$$

Therefore $p$ does not divide the left side. □

We turn now to polynomials in two complex variables $z, w$, which we consider as polynomials in $w$ with coefficients from the ring $C[z]$:

$$P(z,w) \;=\; q_n(z)\, w^n + q_{n-1}(z)\, w^{n-1} + \cdots + q_0(z). \tag{7.4.5}$$

If $q_n$ is not identically zero, we write $\deg_w(P) = n$.

As noted earlier, a polynomial $P(z,w)$ is said to be *irreducible* if it is not a product of non-constant polynomials. If we view $P$ as a polynomial in $w$ with coefficients in the ring $\mathbb{C}[z]$, this means that $P$ is irreducible over $\mathbb{C}[z]$. We can also ask whether it can be factored over the field of rational function $\mathbb{C}(z)$: is $P = P_1 P_2$, where the $P_j$ are non-constant polynomials in $w$ whose coefficients are rational functions of $z$? If not, $P$ is said to be *irreducible over* $\mathbb{C}(z)$. Clearly irreducibility over $\mathbb{C}(z)$ implies irreducibility over $\mathbb{C}[z]$. (A helpful analogy here concerns an ordinary polynomial $p$ with integer coefficients—factoring $p$ as a product of polynomials with integer coefficients, or as a product of polynomials with rational coefficients.) The major result of this section is the converse: irreducibility over $\mathbb{C}[z]$ implies irreducibility over $\mathbb{C}(z)$.

A polynomial of the form (7.4.5) is said to be *primitive* if the coefficients $q_k$ have no common factor. Note that an irreducible polynomial is necessarily primitive.

**Lemma 7.4.3.** *Any polynomial of the form (7.4.5) is a product $a(z)Q(z,w)$ where $a$ is a polynomial and $Q$ is primitive. The factors $a$ and $Q$ are unique up to a constant factor and its inverse.*

*Proof:* Clearly $a$ must be a greatest common divisor of the coefficients $q_k$. This is unique up to a constant factor. □

If the product of two polynomials is primitive, the factors must be primitive. The converse is true as well.

**Lemma 7.4.4.** *The product of primitive polynomials is primitive.*

*Proof:* Suppose that $P$ and $Q$ are primitive, where

$$P = a_0 + a_1 w + a_2 w^2 + \dots, \quad Q = b_0 + b_1 w + b_2 w^2 + \dots .$$

Let $r(z)$ be an irreducible polynomial. Since $P$ and $Q$ are primitive, there are smallest indices $j$ and $k$ such that $r$ does not divide $a_j$ and $b_k$. The coefficient of $w^{j+k}$ in $PQ$ is

$$[a_0 b_{j+k} + \dots + a_{j-1} b_{k+1}] + a_j b_k + [a_{j+1} b_{k-1} + \dots + a_{j+k} b_0].$$

By assumption, $r$ divides each sum in braces. By Lemma 7.4.2, it does not divide $a_j b_k$. Thus there is a coefficient of $PQ$ that is not divisible by $r$. This holds for each such $r$, so $PQ$ is primitive. □

**Lemma 7.4.5.** *Suppose that $P$ is a polynomial in $w$ whose coefficients belong to $\mathbb{C}(z)$. Then there is a rational function $r \in \mathbb{C}(z)$ such that $Q = rP$ is a primitive polynomial.*

*Proof:* Let $b$ be product of the denominators of the coefficients of $P$. Then $bP$ is a polynomial in $w$ with coefficients in $\mathbb{C}[z]$, and Lemma 7.4.3 implies that $bP = aQ$, with $a$ a polynomial and $Q$ primitive. Then $Q = (b/a)P$. □

**Theorem 7.4.6.** *A polynomial $P(z,w)$ is irreducible over $\mathbb{C}[z]$ if and only if it is irreducible over $\mathbb{C}(z)$.*

*Proof:* We may assume that $P$ is primitive. Suppose $P = QR$, with $Q$ and $R$ in $\mathbb{C}(z)$. Choose $a,b,c,d$, polynomials in $z$, such that $(b/a)Q$ and $(d/c)R$ belong to $\mathbb{C}[z]$ and are primitive. Then

$$bd\,P = ac\left[\frac{b}{a}Q \cdot \frac{d}{c}R\right]. \tag{7.4.6}$$

By Lemma 7.4.4, the product in brackets is primitive. Since $P$ is also primitive, it follows from Lemma 7.4.3 that $ac = \lambda bd$ where $\lambda$ is a non-zero constant. Therefore (7.4.6) is equivalent to the factorization over $\mathbb{C}[z]$

$$P = \left(\lambda \frac{b}{a} Q\right)\left(\frac{d}{c} R\right).$$

Since irreducibility over $\mathbb{C}(z)$ implies irreducibility over $\mathbb{C}[z]$, the proof is complete. □

**Corollary 7.4.7.** *Suppose polynomials $P(z,w)$ and $Q(z,w)$ have no common factors over $\mathbb{C}[z]$. Then there are polynomials $R(z,w)$, $S(z,w)$, and $r(z)$ such that $r$ is not identically zero and*

$$R(z,w)\,P(z,w) + S(z,w)\,Q(z,w) \;=\; r(z). \tag{7.4.7}$$

*Proof:* Assume $\deg_w P \geq \deg_w Q$ and consider $P_1(w) = P(z,w)$ and $P_2(w) = Q(z,w)$ as polynomials in $w$ with coefficients in $\mathbb{C}[z] \subset \mathbb{C}(z)$. Apply the division algorithm, leading to

$$P_{m-2} \;=\; C_{m-1}P_{m-1} + P_m,$$

where $P_m$ is independent of $w$ but $P_{m-1}$ is not. If $P_m \equiv 0$, then $P_{m-1}$ divides $P_{m-2}$, as polynomials with coefficients that are rational in $z$, and ultimately $P_{m-1}$ divides $Q$ and $P$. Adapting the proof of Theorem 7.4.6, this leads to a common factor of $P$ and $Q$ over $\mathbb{C}(z)$. Thus $P_m$ is a non-zero rational function of $z$. Working backward, there are polynomials $R_1(z.w)$, $S_1(z.w)$ with coefficients in $\mathbb{C}(z)$ such that

$$R_1\,P + S_1\,Q \;=\; P_m.$$

Multiplying by the product of the denominators of the coefficients of $R_1$, $S_1$, and $P_m$ gives an equation (7.4.7) in polynomials. □

## 7.5 Algebraic curves: some analysis

Consider the function $f(z)$ defined implicitly by the equation $P(z, f(z)) = 0$, where

$$0 \;=\; P(z,w) \;=\; q_n(z)\,w^n + q_{n-1}(z)\,w^{n-1} + \cdots + q_0(z). \tag{7.5.1}$$

We assume that $P$ is irreducible and that $q_n$ is not identically zero. Irreducibility implies that the $q_j$ do not all vanish at any one point $z$. Thus $P(z,\cdot)$ is always a polynomial in $w$ of some degree $\leq n$.

The *critical points* of $P$ are the zeros of the leading coefficient $q_n$, together with the points $z$ for which $P(z,\cdot)$ has a multiple zero $w$. The latter points are the points $z$ such that $P(z,\cdot)$ and the derivative with respect to $w$, $Q(z,\cdot) = P_w(z,\cdot)$, have a common zero $w$. Irreducibility implies that $P$ and $Q$ have no common factor, so Corollary 7.4.7 implies that these latter critical points are among the zeros of a polynomial $r(z)$. Therefore there are finitely many critical points.

**Proposition 7.5.1.** *Suppose that $z_0$ is not a critical point of $P$. Then there is a disk $D_r(z_0)$ and $n$ functions $f_j(z)$ defined and holomorphic in $D_r(z_0)$, such that*

$$P(z, f_j(z)) = 0, \quad \text{if } z \in D_r(z_0);$$
$$f_j(z) - f_k(z') \neq 0, \quad \textit{if } j \neq k \textit{and} z, z' \in D_r(z_0). \tag{7.5.2}$$

*Proof:* By assumption, $P(z_0, \cdot)$ has $n$ distinct zeros $\{w_j(z_0)\}$. Choose $\varepsilon > 0$ small enough that the disks $D_\varepsilon(w_j(z_0))$ are disjoint. Then $P(z_0, w)$ does not vanish for $w$ on any of the circles $\Gamma_j$ that bound these disks. Let $P_w$ denote the partial derivative. Then

$$\frac{1}{2\pi i} \int_{\Gamma_j} \frac{P_w(z_0, w)}{P(z_0, w)} \, dw$$

is the number of zeros of $P(z_0, \cdot)$ in $D_\varepsilon(w_j(z_0))$, i.e. one. For small $r > 0$, $P(z, w)$ will not vanish on any of the $\Gamma_j$ if $|z - z_0| < r$, so $P(z, w)$ will continue to have exactly one zero $w_j(z)$ enclosed by $\Gamma_j$. Moreover the value of this zero is

$$w_j(z) \; = \; \frac{1}{2\pi i} \int_{\Gamma_j} w \, \frac{P_w(z, w)}{P(z, w)} \, dw.$$

The integrand is holomorphic with respect to $z$, so $f_j = w_j$ is holomorphic. □

If a function element $[f_j, D_r(z_0)]$ of Proposition 7.5.1 is continued along a curve, the process will also provide a continuation of $P(z, f_j(z))$, which will continue to vanish identically. Therefore any continuation continues to satisfy $P(z, f(z)) = 0$. In particular, continuation of any $f_j$ around a closed curve that avoids the critical points leads to one of the $f_k$.

We turn to an examination of the $f_j$ near a critical point.

**Proposition 7.5.2.** *Suppose that $z_0 \in \mathbb{C}$ is a critical point or the point at $\infty$, and suppose that $P(z_0, \cdot)$ has a root $w_0$ of multiplicity $k$. Then $z_0$ is the base point for regular points and algebraic branch points of $R_f$, the sum of whose multiplicities is $k$.*

*Proof:* Choose $\rho > 0$ so that $P(z_0, w)$ has no zeros other than $w_0$ in the closed disk $\overline{D}_\rho(z_0)$ and also so that there are no other critical points in this disk. Then at each point of the punctured disk $D_\rho \setminus \{z_0\}$ there are $k$ distinct $f_j$. Continuing each $f_j$ throughout the punctured disk leads to a partition of the set of such $f_j$ into subsets, each of which is closed under analytic continuation in the punctured disk, and thus corresponds to a distinct regular point or algebraic branch point of $R_f$. The sum of the multiplicities is the number of elements of the set that is being partitioned. □

We still need to consider the possibility that, at a critical point, some roots may converge to $\infty$, and also to consider the behavior of roots as $z \to \infty$.

So long as $w \neq 0$, the equation $P(z, w) = 0$ is equivalent to

$$0 \; = \; Q(z, v) \; = \; q_0(z) \, v^n + q_1(z) \, v^{n-1} + \cdots + q_n(z), \qquad v = \frac{1}{w} \neq 0.$$

Now $Q(z_0,0)=0$ if and only if $q_n(z_0)=0$. If so, we can argue as in Proposition 7.5.2 that there are corresponding regular points or branch points with base point $z_0$, that are regular points or algebraic branch points for $v=1/z$ and thus poles or algebraic branch points on the Riemann surface.

Finally, let $N$ be the maximum of the degrees of the $q_k$. Then for $z\neq 0$, $P(z,w)=0$ is equivalent to

$$0 \;=\; R(u,w) \;=\; r_n(u)\,w^n+r_{n-1}(u)\,w^{n-1}+\cdots+r_0(0), \qquad u=\frac{1}{z},$$

where $r_j(u)=u^N q_j(1/u)$. Thus the analysis of $f(1/z)$ for $P(z,f(z))=0$ is qualitatively the same as the analysis of $R(u,g(u))$ for $u$ near 0. We conclude that points of $R_f$ with base point $z=\infty$ also belong to the Riemann surface $R_f$.

**Theorem 7.5.3.** *The Riemann surfaces for the functions $f_j$ coincide: the functions $f_j$ are branches of the analytic function that is determined by the equation $P(z,f(z))=0$.*

*Proof:* Consider the Riemann surface for the continuation of $f=f_1$ in some disk centered at a non-critical point. Each branch of $f$ is one of the $f_j$; we want to show that *each* of the $f_j$ occurs among the continuations of $f_1$. Suppose that exactly $m$ of the $f_j$ occur. Propositions 7.5.1 and 7.5.2, together with the discussion that follows these propositions of the limits $w\to\infty$ and $z\to\infty$, show that the surface is compact. Theorem 7.3.1 shows that there is a polynomial $P_1(z,w)$ of degree $m$ in $w$ such that $P_1(z,f(z))\equiv 0$. There is an irreducible polynomial $Q(z,w)$ of degree $\leq m$ in $w$ such that $Q$ divides $P_1$. If $m<n$ then $P$ and $Q$ have no common factors, and Corollary 7.4.7 applies: there are $R(z,w)$, $S(z,w)$, $r(z)$ such that $r$ is not identically zero and

$$R(z,w)\,P(z,w)+S(z,w)\,Q(z,w) \;\equiv\; r(z).$$

Take $w=f(z)$. Then the left side vanishes identically, a contradiction. Therefore $m=n$. □

## 7.6 Examples: elliptic and hyperelliptic curves

In Section 7.2 we discussed the Riemann surface of the function $\sqrt{z}$, the solution of $w^2-z=0$. A similar approach can be used for any case in which the equation is of second degree in $w$:

$$q_2(z)w^2+q_1(z)w+q_0(z) \;=\; 0, \tag{7.6.1}$$

where the $q_j$ are polynomials and $q_2$ is not identically zero.

**Proposition 7.6.1.** *Equation (7.6.1) is equivalent to an equation*

$$u^2 \;=\; p(z), \tag{7.6.2}$$

*with $u = a_1(z)w + a_2(z)$, where $a_1$ and $a_2$ are certain rational functions, and $p$ is a non-zero polynomial with simple roots.*

For the proof see Exercise 5. Note that the polynomial

$$P(u,z) \;=\; u^2 - p(z)$$

is irreducible; see Exercise 3.

The crucial datum in (7.6.2) is the degree of $p$. For degree 1, the previous construction generalizes easily: slit the Riemann sphere from the zero of $p$ to the point at $\infty$ and join two copies of the slit sphere, along the slit, to make, topologically, a single sphere on which the function $\sqrt{p(z)}$ is single-valued and meromorphic. The same construction works for degree two as well: slit the sphere from one zero of $p(z)$ to the other, and join two copies across the slit. The resulting surface is topologically a sphere; see Figure 7.2.

The same ideas carry over to the general case of (7.6.2), but the topology becomes more interesting. Introduce disjoint slits in the sphere that join disjoint pairs of zeros of $p$ if $p$ has even degree. If $p$ has odd degree, run one slit from one root to $\infty$. A little thought will show that a single-valued branch of $\sqrt{p(z)}$ can be chosen on the slit plane—what needs to be checked is what happens when the function is followed along a curve that encloses a slit. Again, two copies of the slit sphere can be joined along corresponding slits to form a single surface.

The cases $\deg p = 3$, $\deg p = 4$ of this construction involve two slits, and it is easily seen that the resulting surface is a torus; see Figure 7.3. Both are referred to as *elliptic* curves. The torus can be thought of as a sphere with a single hole pushed through it.

In general the resulting surface has $g$ holes, where $g+1 = (\deg p + 1)/2$ if $\deg p$ is odd, $g+1 = \deg p/2$ if $\deg p$ is even; $g$ is called the *genus* of the curve $w^2 = p(z)$. When $g > 1$, the curve is called a *hyperelliptic curve*. The fact that the genus is the same for degree $2m$ and degree $2m-1$ has an algebraic explanation; see Exercise 6.

One thing one might like to do is to find a good parametrization of a given curve $\{(z,w)\}$: a pair of functions $s_1$, $s_2$ of a complex variable $t$ such that

$$s_2(t)^2 \;=\; p(s_1(t)).$$

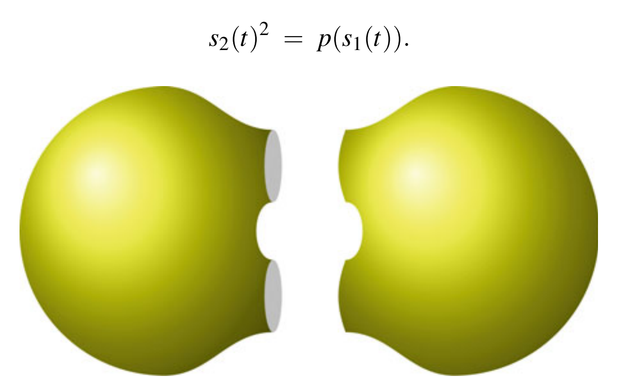


**Fig. 7.3** Joining across two slits

If we impose the additional condition that $s_2$ be the derivative $s_1'$ and set $s = s_1$, then we are considering the equation $s' = \sqrt{p(s)}$. Equivalently

$$\frac{dt}{ds} = \frac{1}{\sqrt{p(s)}}.$$

Then $s(t)$ can be defined implicitly as a function of $t$ by

$$t = \int_{s(0)}^{s(t)} \frac{d\tau}{\sqrt{p(\tau)}}. \tag{7.6.3}$$

When $p$ has degree 1 we may translate and dilate $z$ so that the equation has the form $w = z^2$, and the curve can be parametrized by

$$(z, w) = \left(\frac{t}{2}, \frac{t^2}{4}\right). \tag{7.6.4}$$

When $p$ has degree two, we may translate and dilate $z$ so that the roots of $p$ are $\pm 1$ and the equation is $w^2 = z^2 - 1$. The curve can be parametrized by $(r, \theta)$

$$(z, w) = (\sinh t, \cosh t), \quad t = re^{i\theta}. \tag{7.6.5}$$

Here the parameter space, for finite values of $t$, is an infinite cylinder. Note that both parametrizations (7.6.4) and (7.6.5) are of the form (7.6.3).

For the elliptic case, degrees 3 and 4, see Chapters 16 and 15, respectively.

## 7.7 General compact Riemann surfaces

In this section we sketch the relation between the previous results and a more general point of view.

The general concept of a compact Riemann surface is a complex manifold of one (complex) dimension: a connected topological space $R$ that is covered by finitely many open sets $\Omega_j$, for each of which there is a homeomorphism $\pi_j$ from $\Omega_j$ into $\mathbb{C}$, such that each $\pi_j \circ \pi_k^{-1}$ is holomorphic where defined. A function $g : R \to \mathbb{C}$ is said to be holomorphic (meromorphic) if each $g \circ \pi_j^{-1}$ is holomorphic (meromorphic). The only holomorphic functions defined on all of $R$ are constant (Exercise 4). The meromorphic functions form a field $\mathscr{F}(R)$. This field has been extensively studied, as have certain spaces of differential forms. As we shall see, $\mathscr{F}(R)$ essentially characterizes $R$. We assume here the non-trivial fact that if $R$ is a (non-empty) compact Riemann surface, then $\mathscr{F}(R)$ contains functions that are not constant.

**Lemma 7.7.1.** *If $g$ is a non-constant meromorphic function on the compact Riemann surface $R$, then it takes each value the same number of times (counting multiplicity).*

*Proof:* Recall the proof when $R$ is the Riemann sphere $\mathbb{S}$: changing coordinates by a linear fractional transformation, if necessary, we may assume that $g$ is holomorphic and non-zero at $z = \infty$. Integrating $g'/g$ around a contour $|z| = M$ that encloses all the poles and zeros, we find that the number of poles equals the number of zeros. Replacing $g$ by $g - a$, we see that the number of times that $g$ takes the value $a$ is independent of $a$. This argument carries over to the general situation, if we replace the standard Cauchy integral theorem by Stokes's theorem from differential topology. □

By looking again at local coordinate systems, we may define what it means for a map from one Riemann surface to another to be holomorphic. Two such surfaces $R_1$ and $R_2$ are said to be *biholomorphically equivalent* if there is an invertible holomorphic map $f$ from one to the other. (Note that $f^{-1}$ is necessarily holomorphic if $f$ is.)

**Theorem 7.7.2.** *A (general) compact Riemann surface R is biholomorphically equivalent to the Riemann surface of the function defined by an irreducible polynomial.*

*Proof:* By Lemma 7.7.1, a meromorphic function $u : R \to \mathbb{S}$ takes each value $m$ times, $m = m(u)$. Choose $u$ so that $m(u)$ is minimal. Let $\{z_j\}_{j=1}^k$ be the (finitely many) points in $\mathbb{C}$ that are the images of points that are multiple points of $u$. Choose an additional point $z_0$, and take non-overlapping curves $\gamma_j$ that join $z_0$ to $z_j$. Removing these curves from $\mathbb{S}$ leaves a simply connected domain $\Omega$, whose closure is all of $\mathbb{S}$. The analogue of Theorem 1.7.2 shows that a choice of $u^{-1}$ in a small disk in $\Omega$ has a unique extension to all of $\Omega$. This allows us to find disjoint domains $\widetilde{\Omega}_j$, $1 \le j \le m$ such that $u : \widetilde{\Omega}_j \to \Omega_j$, a copy of $\Omega$, is biholomorphic. Moreover, the closure of the union of the $\widetilde{\Omega}_j$ is all of $R$. Join two of the $\Omega_j$ along a cut if $u^{-1}$ continues from one to the other along that cut. The result, equipped with appropriate coordinate maps, is a surface $R_1$ that is biholomorphic to $R$.

Suppose now that $g$ is a second meromorphic function on $R$ that takes each value $m$ times, and $g$ is not a constant multiple of $u$. Let $f(z) = g \circ u^{-1}(z)$, $z \in \Omega_1$. Continuation of $u^{-1}$ leads to continuation of $f$, and we can conclude that $R$ is the Riemann surface of $f$. It follows from this and the previous results that $R$ is biholomorphically equivalent to the Riemann surface $R_f$ of an algebraic curve. □

## 7.8 Algebraic curves of higher genus

In this section we sketch some of the theory of compact Riemann surfaces of genus $g \ge 2$, with illustrations from genus 2. The first observation is that every compact Riemann surface has a genus, i.e. topologically it is a sphere, a torus, or a surface like that of a hyperelliptic curve—a sphere with two or more holes pushed through. As shown in the previous section, each such surface $R$ can be taken to an algebraic

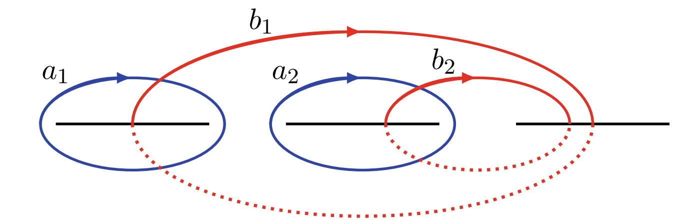



**Fig. 7.4** Cross-cuts in the surface $R$

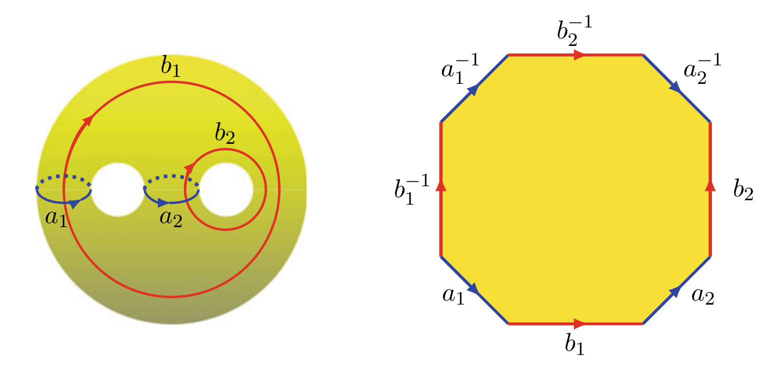



**Fig. 7.5** The surface $R$ with cross-cuts, and the surface after cutting

curve, and therefore viewed as a collection of copies of the sphere $\mathbb{S}$ joined across certain slits.

Consider, for example, the surface for

$$w^2 = P(z),$$

where $P$ has degree 6 and the roots are distinct. Three slits joining pairs of roots of $P$ are shown schematically in Figure 7.4. Figure 7.4 also shows two curves $a_1$ and $a_2$ in the "upper" of two sheets, and two curves $b_1$ and $b_2$ that cross from the upper to the lower sheet through two of the slits.

The surface $R$, formed by joining across the three slits, is depicted on the left in Figure 7.5. The curves $a_j$, $b_j$ are also indicated. Cutting $R$ along these curves results in a simply connected region that is pictured schematically on the right in Figure 7.5.

Reversing the procedure, we may connect the sides $a_j, b_j$ in the right part of Figure 7.5 to their counterparts $a_j^{-1}$, $b_j^{-1}$; the resulting figure is the left part of Figure 7.5. The procedure that we have just outlined works for a general algebraic curve. A judicious choice of cuts yields a simply connected region like that on the right in Figure 7.5, with sides that are paired in a natural way to produce a sphere, a torus, a surface of higher genus like Figure 7.5, or a similar surface with more holes.

As noted in Section 7.6, and proved in the chapters on elliptic functions, elliptic curves (genus 1) can be parametrized by meromorphic functions. As proved in Section 9.4 (in the hyperelliptic case), this is not true for curves of genus $\geq 2$. General Riemann surfaces of genus $\geq 2$ can, however, be parametrized by *automorphic functions*. This is the culmination of nearly a century of work by Abel, Jacobi, Riemann, Klein, Poincaré, Weyl, and others. A complete account of this subject normally takes several hundred pages. Here we outline part of the theory in the special case of a hyperelliptic curve of genus 2. (In fact every compact Riemann surface of genus 2 is equivalent to a hyperelliptic curve; see Farkas and Kra [44], III.7.2.) We need to invoke the notion of homotopy of curves (the possibility of continuous deformation within a given topological space), and the notion of a homotopy group.

Suppose again that $P$ is a polynomial of degree 6 with simple zeros, joined pairwise by slits, as shown schematically in Figure 7.4. (Or suppose that $P$ has degree 5, with simple zeros and take the point at $\infty$ as a sixth point.) Let $\Omega_+$ be the sphere $\mathbb{S}$ with these slits removed. Recall that the Riemann surface $R$ of the curve $w^2 = P(z)$ is obtained by analytic continuation of $w$ along curves beginning at some point $z_0$ in $\Omega_+$. We view passing through one of the slits as taking us to a second copy $\Omega_-$ of the slit sphere, and crossing a slit in $\Omega_-$ brings us back to $\Omega_+$.

The same kind of construction gives rise to the *covering manifold* $\widetilde{R}$ of the Riemann surface. Here we continue the coordinate $z$ itself along curves starting from $z_0 \in \Omega_+$ as a function $\widetilde{w} = z$. However the continuations along curves $\gamma_1$ and $\gamma_2$ are considered as giving *distinct values*, unless $\gamma_1$ and $\gamma_2$ are homotopic in $R$. This means, for example, that crossing one of the cuts indicated in Figure 7.4 and back through a second cut leads not to the original manifold, but to a copy that may be indexed by the two crossings (taking into account the order of the crossings). This process is continued along curves that make any (finite) number of crossings, leading to a manifold with infinitely many sheets.

The resulting manifold $\widetilde{R}$ is a covering manifold: there is a map $\pi : \widetilde{R} \to R$ with the property that for each point $p \in R$, there is a neighborhood $U$ of $p$ such that $\pi^{-1}(U)$ consists of disjoint copies of $U$. Moreover, the surface $\widetilde{R}$ is simply connected. This follows by tracking down the definition: if two curves from 0 arrive at points $p_1, p_2$ with the same projection $\pi(p) \in R$, then the curves are homotopic if and only if the points $p_j$ lie on the same branch of $\widetilde{R}$. The manifold $\widetilde{R}$ itself has a complex structure, induced locally by the projection map $\pi$.

The next stage of the development is to observe that, associated with this construction, there is a natural group of bijective conformal transformations $\{A\}$ of $\widetilde{R}$, with the property that $\pi \circ A = \pi$. Such maps are called *covering transformations*, or *deck transformations*.

In fact, suppose that $p_0$ is a point of $R$ and that $\gamma$ is a closed curve in $R$ with $p_0$ as starting point. Suppose that $\widetilde{p}_0$ belongs to $\pi^{-1}(p_0)$. The curve $\gamma$ lifts via $\pi^{-1}$ to a curve $\widetilde{\gamma}$ in $\widetilde{R}$ with starting point $\widetilde{p}_0$. The end point of $\widetilde{\gamma}$ also lies in $\pi^{-1}(p_0)$; we define it to be $A_\gamma(\widetilde{p}_0)$. This map can be extended by continuity to a neighborhood. In fact choose a small (connected, simply connected) neighborhood $U$ of $p_0$ such that $\pi^{-1}(U)$ consists of disjoint copies of $U$. For each $p \in U$ there is a curve homotopic

to $\gamma$ that has starting point $p$, and the resulting action takes the point in $\pi^{-1}(p)$ that lies near $\widetilde{p}_0$ to the point in $\pi^{-1}(p)$ that lies near $A_\gamma(\widetilde{p}_0)$. Following this argument through a covering of $R$ by such neighborhoods $U$, we extend $A_\gamma$ to all of $\widetilde{R}$. The map $A_\gamma$ depends only on the homotopy class of $\gamma$. Moreover $A_\gamma$ has no fixed points unless $\gamma$ is homotopic to a constant map; see Exercise 10.

It can be shown that the map from $\gamma$ to $A_\gamma$ in the group $\mathrm{Aut}(\widetilde{R})$ of covering transformations is an isomorphism from the homotopy group $H_{1,p_0}(R)$ of closed curves based at $p_0$ to the group $G = \mathrm{Aut}(\widetilde{R})$. A result of all this is that $R$ can be identified with the quotient $\widetilde{R}/G$, constructed by identifying two points if and only if some $A_\gamma$ carries one to the other.

We equipped $\widetilde{R}$ with the complex structure that is given locally by using $\pi^{-1}$ to pull back the complex structure on $R$. Then $\widetilde{R}$ is a simply connected Riemann surface. This makes possible a different description/construction of $R$. It is a (not trivial) fact that each simply connected Riemann surface is conformally equivalent to either the complex plane $\mathbb{C}$, the Riemann sphere $\mathbb{S}$, or the unit disk $\mathbb{D}$. It is also a fact that for surfaces of genus $g \geq 2$, the covering manifold is conformally equivalent to the disk. Therefore the previous discussion leads to the conclusion that each compact Riemann surface of genus $\geq 2$ can be identified with the quotient

$$\mathrm{Aut}(\mathbb{D})/\Gamma,$$

where $\Gamma$ is a subgroup of the automorphism group of the unit disk $\mathbb{D}$ (which consists of certain linear fractional transformations: Proposition 2.3.2). The non-identity elements of $\Gamma$ are fixed-point free. Moreover, $\Gamma$ is discrete: each element of $\Gamma$ has a neighborhood containing no other element of $\Gamma$; see Exercise 11.

This result is known as the *uniformization theorem*. A consequence is that each Riemann surface of genus $g \geq 2$ can be parametrized by *automorphic functions*: functions on $\mathbb{D}$ that are invariant under the discrete subgroup $\Gamma$; see Exercise 12.

**Remarks.** The uniformization theorem is a non-constructive theorem. Starting from an algebraic curve, there is no algorithm to determine the associated group $G$ of covering transformations, or the automorphic functions to parametrize the curve. Thus the following example is somewhat exceptional.

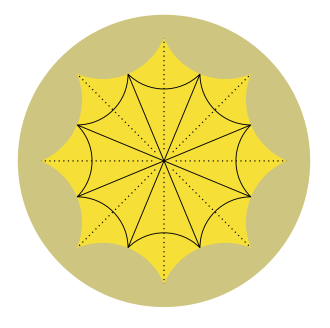
An example in genus 2 is the *Bolza curve*, also called the Bolza surface. This is the curve

$$w^2 \;=\; P(z) \;=\; z(z^2-1)(z^2+1). \tag{7.8.1}$$

The automorphism group $\Gamma \subset \mathrm{Aut}(\mathbb{D})$ can be taken to have generators

$$A_k \;=\; \begin{bmatrix} 1+\sqrt{2} & (2+\sqrt{2})\sqrt{\sqrt{2}-1}\,\exp(ik\pi/4) \\ (2+\sqrt{2})\sqrt{\sqrt{2}-1}\,\exp(-ik\pi/4) & 1+\sqrt{2} \end{bmatrix},$$

$k = 0,1,2,3$. A *fundamental domain* for $\Gamma$ is a minimal domain with the property that the images under $\Gamma$ of its closure fill out $\mathbb{D}$. The highlighted area $\Omega$ in the image of $\mathbb{D}$ on the left in Figure 7.6 is such a domain. It is a curvilinear polygon, in fact a



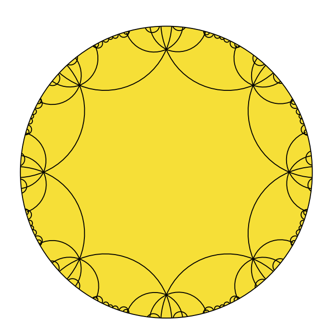


**Fig. 7.6** Fundamental domain $\Omega \subset \mathbb{D}$

regular curvilinear octagon partitioned into curvilinear triangles with vertex angles $\pi/4$, see Sections 6.4, 6.5, 6.6. The sides of each triangle are portions of geodesics with respect to the hyperbolic metric in $\mathbb{D}$, and the elements of $\Gamma$ are isometries with respect to this metric; see Chapter 3. Some of the images of the fundamental domain $\Omega$ can be seen in the image of $\mathbb{D}$ on the right in Figure 7.6. (Successive images get smaller very rapidly, with respect to the euclidean distance, as one approaches the boundary of $\mathbb{D}$.)

The actions of $\Gamma$ can be visualized here. They are generated by

(a) rotations by $\pi/4$;

(b) rotation of order 3 around the center of any of the 16 small triangles bounded by solid lines, or a solid line and the boundary of $\Omega$;

(c) reflection around any of the dotted lines (extended to the boundary of $\mathbb{D}$);

(d) reflection through a side of any of the 16 small triangles.

In fact $\Gamma$ consists of all products of these that contain an even number of reflections. As was the case for the octagon on the right in Figure 7.5, appropriate sides of the octagon in Figure 7.6 may be identified under these actions. This leads to a model of the curve like that on the left in Figure 7.5. The dimensions of $\Omega$ can be deduced from these properties, Exercise 13.

**Remarks:** The reason for the nomenclature here is obscure. The equation (7.8.1) does not occur explicitly in Bolza's paper [26], which is devoted to equations of degree six. The first treatment of (7.8.1) seems to be due to Burnside [29], who determined an explicit representation of the curve using rational functions of the Weierstrass $\wp$ function and its derivatives. Burnside also determined the associated Fuchsian equation; for this connection, see Sections 6.4–6.6. For more on (7.8.1), see [28]. For uniformization of general surfaces of genus 2 see [81].

## Exercises

1. Determine the branches of the Riemann surface for $\sin^{-1} z$ with base point 1.
2. Determine the location and type of the singular points for the functions $w = f(z)$ defined implicitly by each of the following equations. Can you determine the topological nature of the associated Riemann surface?

   (a) $w^3 - 3wz + 2z^3 = 0$.

   (b) $w(w^2 - z^2) - z^4 = 0$.

   (c) $w^2 = (z^2 - 1)(z^2 - 4)$.
3. Suppose that the polynomial $P(z)$, of positive degree, has no multiple roots. Prove that $Q(z,w) = w^2 - P(z)$ is irreducible.
4. Prove that a holomorphic function $f : R \to \mathbb{C}$ on a (general) compact Riemann surface is necessarily constant.
5. Suppose that $r_1$ and $r_0$ are rational functions. Show that the following sequence of equations are equivalent to

$$w_0^2 \;=\; r_1(z)w_0 + r_0(z).$$

   In each case, $w_{j+1} = a_j(z)w_j + b_j(z)$, where $a_j$ and $b_j$ are rational functions of $z$ and $a_j \equiv 1$ if $b_j \neq 0$, while the $p_j$, $q_j$ are polynomials in $z$.

   (a) $w_1^2 = p_1/q_1$.

   (b) $w_2^2 = p_1 q_1$.

   (c) $w_3^2 = p_2$, where $p_2$ has only simple roots.
6. Suppose that $p_2$ in part (c) of Exercise 5 has even degree $2m$. Replace $z$ by a certain linear fractional transformation $z'(z)$ and show that the equation (c) is equivalent to $w^2 = p(z')$, where $p$ has degree $2m - 1$.
7. Discuss (7.6.3) in the case when $p(s) \;=\; 1 - s^2$ and $s(0) = 0$.
8. With the assumptions and notation of Theorem 7.7.2, show that a consequence of the theorem is that $g$ satisfies an equation

$$g^n + r_{n-1}(u)\,g^{n-1} + \cdots + r_1(u)\,g + r_0(u) \;=\; 0, \tag{7.8.2}$$

   where the $r_j$ are rational functions.
9. Give a direct proof of (7.8.2).
10. Prove that a non-identity covering transformation has no fixed points.
11. Show that each element of $\Gamma$ is isolated in $\Gamma$.
12. Show that the Bolza curve, in the form $\mathrm{Aut}(\mathbb{D})/\Gamma$, can be parametrized by automorphic functions.
13. (a) As noted above, each of the triangles with vertex at the center in Figure 7.6 can be mapped to itself by an element of $\mathrm{Aut}(\mathbb{D})$ that rotates the vertices. Use this fact to determine the length of the straight side.

    (b) Reflection over the curved side of the previous triangle maps to the smaller triangle. Use this fact to determine the length of the dotted line.

## Remarks and further reading

We have barely touched on the general subject of compact Riemann surfaces or general Riemann surfaces. There are many modern treatments of varying depth, e.g. Cavalieri [31], Donaldson [36], Miranda [99], Narasimhan [102], Schlag [125]. Farkas and Kra [44], [45] is particularly comprehensive.

For an efficient treatment of covering spaces and the basics of automorphic function theory, see Siegel [127], [128], Chapters 2 and 3. The classic work is Weyl [140].

For more on the algebraic theory, see Goldschmidt [51]. For algebraic curves over finite fields, and applications, see Ling, Wang, and Xing [91].

The uniformization theorem was first stated by Klein and by Poincaré, and eventually proved by Koebe and Poincaré. Abikoff [1] has a very readable exposition of the uniformization theorem, its history, and related issues. A modern proof, based on quasiconformal mapping, is due to Bers [21]. For some different perspectives, see Chapter 1 of Hubbard [68]. A full account of the history is given in Gray [53].

# Chapter 8
# Entire functions




An entire function, a function that is defined and holomorphic in the entire plane $\mathbb{C}$, can be analyzed in terms of its zeros and of its growth. Such an analysis has important applications.

Consider the question of the possible zeros of a non-constant entire function $f$. It can have no zeros, like $e^z$, or a given finite collection of zeros, like a polynomial with precisely those roots. There are only finitely many zeros in each disk $D_R(0)$, so the number of zeros is at most countable. This is all that can be said at this level of generality: the product theorem of Weierstrass shows that each sequence $\{z_n\}_{n=1}^{\infty}$ in $\mathbb{C}$ such that $|z_n| \to \infty$ is the set of zeros of an entire function.

Conditions on the growth of $f$ at infinity have a bearing on the possible distribution of zeros, and conversely, as shown by Jensen's inequality and by a theorem of Hadamard. One application of Hadamard's theorem is to Riemann's xi function of Section 11.5. The xi function plays a key role in the analysis of the Riemann zeta function and its relation to the distribution of primes; see Chapter 13.

The chapter concludes with a simple application of Hadamard's factorization theorem to an eigenvalue problem.

## 8.1 The Weierstrass product theorem

The key ingredients in the proof of the product theorem are factors of the form

$$E(w,p) \;=\; (1-w)\exp\left\{w+\frac{w^2}{2}+\cdots+\frac{w^p}{p}\right\}.$$

The basic fact about such a factor is the estimate

$$\log|E(w,p)| \;\le\; \frac{2|w|^{p+1}}{p+1} \;\text{ if }\; |w| \;\le\; \frac{1}{2}. \tag{8.1.1}$$



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_8

In fact $|w| < 1$ implies

$$\log(1-w) = -\sum_{n=1}^{\infty} \frac{w^n}{n}$$
$$= -\sum_{n=1}^{p} \frac{w^n}{n} - \frac{w^{p+1}}{p+1}\left[1 + \frac{(p+1)w}{p+2} + \frac{(p+1)w^2}{p+3} + \dots\right].$$

The term in brackets is dominated by $\sum_0^\infty |w|^k = 1/(1-|w|)$, if $|z| \le 1/2$, which confirms (8.1.1). This estimate leads to the Weierstrass product theorem.

**Theorem 8.1.1.** (Weierstrass) *Let $z_n$ be a sequence of complex numbers such that $|z_n| \le |z_{n+1}|$ and $|z_n| \to \infty$ as $n \to \infty$. There is an entire function $f$ whose zeros are precisely the $z_n$ (counting multiplicity).*

*Proof:* For convenience, we may assume that 0 is not among the $z_n$. Let $r_n = |z_n|$. The idea is to let

$$f(z) = \prod_{n=1}^{\infty} E\left(\frac{z}{z_n}, p_n\right) = \prod_{n=1}^{\infty}\left(1 - \frac{z}{z_n}\right)\exp\left\{\frac{z}{z_n} + \cdots + \frac{1}{p_n}\left(\frac{z}{z_n}\right)^{p_n}\right\}, \quad (8.1.2)$$

for some choice of integers $\{p_n\}$. If this choice can be made so that the product converges uniformly on each bounded set, then $f$ has the desired properties.

Suppose that $|z| \le r$. Then for $n$ such that $r_n \ge 2r$ and $N > n$, (8.1.1) gives

$$\left|\log \prod_{k=n}^{N} E\left(\frac{z}{z_k}, p_k\right)\right| \le \sum_{k=n}^{N} \frac{2}{p_k+1}\left(\frac{r}{r_k}\right)^{p_k+1}.$$

Therefore uniform convergence follows if we choose the $p_k$ such that for each fixed $r > 0$,

$$\sum_{n=1}^{\infty} \frac{1}{p_n+1}\left(\frac{r}{r_n}\right)^{p_n+1} < \infty. \quad (8.1.3)$$

The choice $p_n = n-1$ will work in all cases, since only finitely many $r_n$ are $\le 2r$. (This choice is far from best possible—see Exercises 1 and 2.) □

**Corollary 8.1.2.** *If $f$ is a non-zero entire function, it has a factorization*

$$f(z) = z^d P(z) e^{h(z)}, \quad (8.1.4)$$

*where $P$ is a (finite or infinite, possibly empty) product of the form (8.1.2) and $h$ is an entire function.*

*Proof:* Choose $d$ so that $z^{-d} f(z)$ is entire and non-zero at the origin. Let $\{z_n\}$ be the zeros of this new function and let $P$ be a polynomial, or a convergent Weierstrass product, with these zeros. Then the function

$$g(z) = z^{-d}\frac{f(z)}{P(z)}$$

is entire and has no zeros. Therefore we may choose a branch of the logarithm $h(z) = \log g(z)$, and this function is entire. □

## 8.2 Jensen's formula

The main result of this chapter is a factorization theorem of Hadamard that gives much more information about a certain class of entire functions.

An important step is to relate the number of zeros of an entire function to its rate of growth. The key result is *Jensen's formula*. This formula gives a relation between the zeros of an entire function $f$ and the growth of $|f|$. It relies on the mean value property, which we repeat here. If $f$ is holomorphic for $|z| < r$ and continuous on the closure, then

$$\operatorname{Re} f(0) = \frac{1}{2\pi}\int_0^{2\pi} \operatorname{Re} f(re^{i\theta})\,d\theta. \tag{8.2.1}$$

**Theorem 8.2.1.** (Jensen) *Suppose that $f$ is entire, $f(0) \neq 0$, and the zeros of $f$, repeated according to multiplicity, are $\{z_n\}_1^\infty$, with $|z_1| \le |z_2| \le |z_3| \le \dots$ . Then*

$$\log\left(\frac{|f(0)|r^n}{|z_1 z_2 \cdots z_n|}\right) = \frac{1}{2\pi}\int_0^{2\pi} \log|f(re^{i\theta})|\,d\theta, \qquad |z_n| \le r < |z_{n+1}|. \tag{8.2.2}$$

*Proof:* By continuity, it is enough to consider $|z_n| < r < |z_{n+1}|$. We know from Proposition 2.3.2 that the linear fractional transformation

$$f_k(w) = \frac{w - w_k}{\bar{w}_k w - 1}, \qquad w_k = \frac{z_k}{r} \tag{8.2.3}$$

has modulus 1 for $|w| = 1$. Therefore $g_k(z) = f_k(z/r)$ has the properties

$$g_k(z_k) = 0, \qquad |g_k(0)| = \frac{|z_k|}{r}, \qquad |g_k(z)| = 1 \text{ if } |z| = r. \tag{8.2.4}$$

The function $g = f/(g_1 g_2 \cdots g_n)$ has no zeros in the disk of radius $r$ centered at the origin, so we may choose a branch of $\log g$ holomorphic in this disk. The real part of $\log g$ is $\log|g|$. Formula (8.2.2) follows from (8.2.3), (8.2.4), and (8.2.1). □

Let $n(s)$ denote the number of $|z_k| \le s$:

$$n(s) = \sup\{k : |z_k| \le s\}; \qquad n(s) = 0 \text{ if } s < |z_1|.$$

Then an equivalent form of Jensen's formula (8.2.2) is

$$\log|f(0)| + \int_0^r \frac{n(s)\,ds}{s} = \frac{1}{2\pi}\int_0^{2\pi} \log|f(re^{i\theta}|\,d\theta; \tag{8.2.5}$$

see Exercise 3.

A natural tool used in the analysis of entire functions is the maximum modulus on circles

$$M(r) = M(f,r) = \sup_{|z|=r} |f(z)|.$$

This is continuous and, by the maximum principle, non-decreasing. Jensen's formula gives an estimate for $n(r)$ in terms of $M$.

**Corollary 8.2.2.** *Under the hypotheses of Theorem 8.2.1, there is a constant C such that*

$$n(r) \leq \frac{\log M(2r)}{\log 2} + C. \tag{8.2.6}$$

*Proof:* The right side of formula (8.2.5) is at most $\log M(r)$. The function $n$ is non-decreasing, so

$$\int_0^{2r} \frac{n(s)\,ds}{s} \geq \int_r^{2r} \frac{n(r)\,ds}{s} = n(r)\log 2.$$

Using (8.2.5) with $2r$ in place of $r$, we obtain (8.2.6) with $C = |\log|f(0)|/\log 2|$. □

The extended form of Liouville's theorem, Theorem 1.2.8, implies that $M(r) = O(r^n)$ as $r \to \infty$, where $n$ is a positive integer, if and only if $f$ is a polynomial of degree $\leq n$. What if $f$ is not a polynomial?

We start here with another look at the Weierstrass product (8.1.2). The optimal case for convergence of the product is when it suffices to take a fixed $p_n = p$, for every $n$: the case when

$$\sum_{n=1}^{\infty} \frac{1}{|z_n|^{p+1}} < \infty. \tag{8.2.7}$$

If this is the case, and if $p$ is the smallest integer that gives convergence, then the product

$$\prod_{n=1}^{\infty} E\left(\frac{z}{z_n}, p\right) \tag{8.2.8}$$

is said to be the *canonical product* for the zeros $\{z_n\}$. This value of $p$ is called the *genus*.

**Proposition 8.2.3.** *If the set of zeros $\{z_n\}$ satisfies (8.2.7) for some integer $p \geq 0$, then the function P defined by the product (8.2.8) satisfies the estimate*

$$|P(z)| \leq \exp(C|z|^{p+1}), \tag{8.2.9}$$

*for some constant C.*

*Proof:* It is easy to see that $|w| \geq 1/2$ implies that

$$|E(w,p)| \le (1+|w|)\exp\{p|2w|^p\} \le 3|w|\exp\{p|2w|^p\}$$
$$\le \exp\left\{(p+1)|2w|^{p+1}\right\},$$

since $3|w| \le \exp(3|w|) \le \exp(|2w|^{p+1})$ if $2|w| \ge 1$. Combining this with the estimate (8.1.1) for $|w| \le 1/2$, we find that each factor of $P$ satisfies the estimate

$$\left|E\left(\frac{z}{z_k}, p\right)\right| \le \exp\left(C_0 \frac{|z|^{p+1}}{|z_k|^{p+1}}\right), \qquad C_0 = (p+1)2^{p+1}.$$

Therefore (8.2.9) holds with $C = C_0 \sum_k (1/|z_k|^{p+1})$. □

This leads us to the concept of finite order.

## 8.3 Functions of finite order

An entire function $f$ is said to be of *finite order* if there are constants $C$ and $\rho$ such that

$$|f(z)| = O\left(e^{C|z|^\rho}\right) \qquad \text{as } |z| \to \infty. \tag{8.3.1}$$

The *order* of $f$ is the greatest lower bound of the $\rho$ for which such an estimate holds. For example, polynomials have order zero. If $Q$ is a polynomial of degree $n$, then $e^{Q(z)}$ has order $n$, in particular, $e^z$ has order 1. The function $e^{e^z}$ does not have finite order.

Proposition 8.2.3 says that a Weierstrass product with genus $p$ has order at most $p+1$. The following is a partial converse. It implies, in particular, that if the product in Proposition 8.2.3 is canonical and has order $\rho$, then $p < \rho$ if $\rho$ is not an integer, $p \le \rho + 1$ if $\rho$ is an integer.

**Proposition 8.3.1.** *Suppose the entire function $f$ has order $\rho$. Let $\{z_n\}$ be the set of non-null zeros of $f$, with $|z_n| \le |z_{n+1}|$, repeated according to multiplicity. Then for each $\sigma > \rho$,*

$$\sum_{n=1}^{\infty} \frac{1}{|z_n|^\sigma} < \infty. \tag{8.3.2}$$

*Proof:* We may replace $f$ by $z^{-d} f$ if necessary, and assume that $f(0) \ne 0$. Combining the order assumption with the estimate (8.2.6), we find that for a given $\varepsilon > 0$,

$$n(r) \le C_\varepsilon r^{\rho+\varepsilon}.$$

Setting $r = |z_n|$ gives $n \le C_\varepsilon |z_n|^{\rho+\varepsilon}$, or

$$\frac{1}{|z_n|^{\rho+\varepsilon}} = O\left(\frac{1}{n}\right). \tag{8.3.3}$$

Given $\sigma > \rho$, choose $\varepsilon$ so that $\rho + \varepsilon < \sigma$. Then (8.3.3) implies

$$\frac{1}{|z_n|^\sigma} = O\left(\frac{1}{n^\tau}\right), \quad \tau = \frac{\sigma}{\rho+\varepsilon} > 1. \qquad \square$$

The final tool needed for the proof of the Hadamard factorization theorem is an estimate for intermediate values of $M(r)$ using the analogous supremum for the real part.

**Theorem 8.3.2.** (Borel, Caratheodory) *Suppose that $f$ is holomorphic in the disk $D_R(0)$ and continuous on the closure. Then for $0 < r < R$,*

$$M(r) \le \frac{2r}{R-r} A(R) + \frac{R+r}{R-r} |f(0)|, \qquad A(R) = \sup_{|z|=R} \operatorname{Re} f(z). \tag{8.3.4}$$

Proof. We may assume that $f$ is not constant. Suppose that $f(0) = 0$, so $A(R) > 0$. Let

$$h(z) = \frac{f(z)}{2A(R) - f(z)},$$

so

$$f(z) = \frac{2A(R)\, h(z)}{1 + h(z)}.$$

Let $f = u + iv$, $u$ and $v$ real-valued. Then

$$|2A(R) - f(z)|^2 = |2A(R) - u(z)|^2 + |v(z)|^2 \ge |u(z)|^2 + |v(z)|^2 = |f(z)|^2,$$

so $|h(z)| \le 1$ for $z$ in the disk. Since $h(0) = 0$, the Schwarz lemma, Lemma 2.3.3, adapted to $D_R(0)$, implies $|h(z)| \le |z|/R$, so

$$|f(z)| \le \frac{2A(R) r/R}{1 - r/R} = \frac{2A(R)\, r}{R-r}. \tag{8.3.5}$$

If $f(0) \ne 0$, we consider $g(z) = f(z) - f(0)$ instead. Then (8.3.5) applied to $g$ gives

$$M(r) - |f(0)| \le \frac{2[A(R) + |f(0)|] r}{R-r}$$

which is (8.3.4). $\square$

**Corollary 8.3.3.** *Under the hypotheses of Theorem 8.3.2,*

$$\sup_{|z|=r} |f^{(n)}(z)| \le \frac{n!\, 2^{n+2} R}{(R-r)^{n+1}} [A(R) + |f(0)|]. \tag{8.3.6}$$

*Proof:* Using Cauchy's formula for the derivative, and integrating over the circle with center $z$ and radius $(R-r)/2$, we find

$$|f^{(n)}(z)| \le n!\left(\frac{2}{R-r}\right)^n M\left(\frac{R+r}{2}\right). \tag{8.3.7}$$

By (8.3.4), since $R - \frac{1}{2}(R+r) = \frac{1}{2}(R-r)$,

$$\begin{aligned} M\left(\frac{R+r}{2}\right) &\le \frac{2(\frac{1}{2}(R+r))}{\frac{1}{2}(R-r)} A(R) + \frac{R+\frac{1}{2}(R+r)}{\frac{1}{2}(R-r)}\,|f(0)| \\ &\le \frac{4R}{R-r}\,[A(R)+|f(0)|]. \end{aligned} \tag{8.3.8}$$

The estimate (8.3.6) follows from (8.3.7) and (8.3.8). □

## 8.4 Hadamard's factorization theorem

We are finally in a position to state and prove the factorization theorem of Hadamard [55].

**Theorem 8.4.1.** *Suppose that $f$ is an entire function of finite order $\rho$. Then*

$$f(z) = z^d P(z)\, e^{Q(z)}, \tag{8.4.1}$$

*where $P$ is the canonical product associated to the non-null zeros of $f$ and $Q$ is a polynomial of degree $n \le \rho$.*

*Proof:* Corollary 8.1.2 and Proposition 8.3.1 imply that $f$ has a factorization of the form (8.4.1), and that the canonical product $P$ has genus $p < \rho + 1$. It remains to be shown that $Q$ is a polynomial of degree $\le n = [\rho]$, the greatest integer $\le \rho$. For convenience, let us replace $f$ by $z^{-d} f$ and assume that $f(0) \ne 0$. We want to show that the derivative $Q^{(n+1)}$ vanishes. The logarithmic derivative of $f$ is

$$\frac{f'(z)}{f(z)} = -\sum_{m=1}^{\infty}\left[\frac{1}{z_m - z} - q_m'(z)\right] + Q'(z),$$

where

$$q_m(z) = \frac{z}{z_m} + \frac{z^2}{2z_m^2} + \cdots + \frac{z^p}{p\, z_m^p}.$$

Now $p+1$ is at most the first integer $> \rho$, so $p \le n$. It follows that differentiating the logarithmic derivative $n$ times gives

$$\left\{\frac{f'(z)}{f(z)}\right\}^{(n)} = -\sum_{m=1}^{\infty} \frac{n!}{(z_m - z)^{n+1}} + Q^{(n+1)}. \tag{8.4.2}$$

Now set

$$g_R(z) = \frac{f(z)}{f(0)} \prod_{|z_m| \le R} \left(1 - \frac{z}{z_m}\right)^{-1}. \tag{8.4.3}$$

Note that $g_R$ is entire and has no zeros with modulus $\le R$, and $g_R(0) = 1$. Let $h_R(z) = \log g_R(z)$, $|z| \le R$, taking the principal branch of the logarithm. Then

$$\begin{aligned} h_R^{(n)}(z) &= \left\{\frac{f'(z)}{f(z)}\right\}^{(n)} + \sum_{|z_m| \le R} \frac{n\,!}{(z_m - z)^{n+1}} \\ &= Q^{(n+1)} - \sum_{|z_m| > R} \frac{n\,!}{(z - z_m)^{n+1}}. \end{aligned} \tag{8.4.4}$$

The final step is to estimate $h_R^{(n+1)}$. If $|z| = 2R$ then each $(1 - z/z_m)^{-1}$ in the product (8.4.3) has modulus $\le 1$, so

$$|g_R(z)| \le \left|\frac{f(z)}{f(0)}\right| \le e^{CR^{\rho+\varepsilon}}.$$

Since $g_R$ is entire, this inequality then also holds for $|z| \le R$. Therefore

$$\mathrm{Re}\,[h_R(z)] = \log |g_R(z)| \le CR^{\rho+\varepsilon}, \qquad |z| \le R.$$

By Corollary 8.3.3, for $|z| = r < R$,

$$|h_R^{(n+1)}(z)| \le \frac{2^{n+3}(n+1)\,!}{(R-r)^{n+2}} CR^{\rho+1+\varepsilon}.$$

In particular, for $|z| \le \frac{1}{2}R$,

$$|h_R^{(n+1)}(z)| \le 2^{2n+5}(n+1)\,!\,CR^{\rho+\varepsilon-(n+1)}.$$

Combining this with (8.4.4) we find that for $|z| < R/2$,

$$|Q^{(n+1)}(z)| \le 2^{2n+5}(n+1)\,!\,CR^{\rho+\varepsilon-(n+1)} + n\,!\,2^{n+1} \sum_{|z_m| > R} \frac{1}{|z_m|^{n+1}}.$$

For small $\varepsilon$ the first term on the right has limit 0 as $R \to \infty$. The sum on the right converges so it also has limit 0 as $R \to \infty$ Therefore the entire function $Q^{(n+1)}$ is identically 0. □

## 8.5 Application to Riemann's xi function

This section depends on the discussion of the xi function

$$\xi(s) = \frac{s(1-s)}{2}\Gamma\left(\frac{s}{2}\right)\pi^{-s/2}\zeta(s) \tag{8.5.1}$$

in Chapter 11.

The original motivation for Hadamard's study of entire functions was to prove Riemann's assertion that the xi function has a factorization

$$\xi(s) = \xi(0)\prod_{\rho}\left(1-\frac{s}{\rho}\right), \tag{8.5.2}$$

where the $\rho$ are the zeros of the entire function $\xi$, which are the same as the non-trivial zeros of the Riemann zeta function $\zeta(s)$. (The "trivial zeros" of $\zeta$ are the negative even integers.) By (11.5.5), $\xi(0) = 1/2$.

For this purpose, we need to estimate $\xi$. The symmetry $\xi(s) = \xi(1-s)$ implies that we may restrict attention to the half plane $\{s : \mathrm{Re}\, s \geq \frac{1}{2}\}$. It is easy to see that for $\mathrm{Re}\, s \geq 2$,

$$|\zeta(s)| \leq \sum_{n=1}^{\infty}\frac{1}{n^{\mathrm{Re}\, s}} \leq \sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}. \tag{8.5.3}$$

For the range $\frac{1}{2} \leq \mathrm{Re}\, s < 2$ we take advantage of (11.1.2): for $\mathrm{Re}\, s > 0$,

$$\Gamma(s)\zeta(s) = \frac{1}{s-1} - \frac{1}{2s} + \int_0^1 f(x)x^{s-1}\,dx + \int_1^{\infty}\frac{x^{s-1}}{e^x-1}\,dx, \tag{8.5.4}$$

where $f(x) = (e^x-1)^{-1} - 1/x + 1/2 = O(x)$ as $x \to 0$. It follows from this identity that

$$\left|\Gamma(s)\zeta(s) - \frac{1}{s-1}\right| \leq \text{constant}, \quad \frac{1}{2} \leq \mathrm{Re}\, s < 2. \tag{8.5.5}$$

Now

$$|\Gamma(s)| \geq \mathrm{Re}\,\Gamma(s) = \int_0^{\infty} e^{-t}\mathrm{Re}\,(t^{s-1})\,dt,$$

which is bounded below for $1/2 \leq \mathrm{Re}\, s \leq 2$. We know that $\zeta$ has a simple pole with residue 1 at $s = 1$, so (8.5.5) implies that

$$\left|\zeta(s) - \frac{1}{s-1}\right| \leq \text{constant}, \quad \frac{1}{2} \leq \mathrm{Re}\, s \leq 2. \tag{8.5.6}$$

It follows from (8.5.1) and (8.5.3)–(8.5.6) that $|\xi(s)|$ is dominated by

$$|s|^2\left[1+\Gamma\left(\frac{s}{2}\right)\right], \quad \mathrm{Re}\, s \geq \frac{1}{2}.$$

In view of the approximation (10.5.8), this implies

$$|\xi(s)| = O(e^{|s|\log|s|}); \tag{8.5.7}$$

see Exercise 7. Thus the entire function $\xi$ has order 1. It also follows from (8.5.1) and (10.5.2) that, as an estimate in terms of $|s|$, (8.5.7) is optimal.

The optimality of (8.5.7) and Proposition 8.2.3 implies that the sum

$$\sum_{\rho} \frac{1}{|\rho|} = \infty.$$

Thus the product (8.5.2) is not absolutely convergent. On the other hand, it follows from Proposition 8.3.1 that

$$\sum_{\rho} \frac{1}{|\rho|^{1+\varepsilon}} < \infty \quad \text{for each } \varepsilon > 0.$$

In particular, $\sum |\rho|^{-2}$ is finite.

**Theorem 8.5.1.** *(Hadamard) The product formula*

$$\xi(s) = \xi(0) \prod_{\xi(\rho)=0} \left(1 - \frac{s}{\rho}\right) \tag{8.5.8}$$

*is valid, in the sense that*

$$\frac{\xi(s)}{\xi(0)} = \lim_{N\to\infty} \prod_{|\rho|\le N} \left(1 - \frac{s}{\rho}\right). \tag{8.5.9}$$

*Proof:* Consider the canonical product for this case. We have shown that $\xi$ has order 1, so the canonical product is

$$\prod_{\rho} \left(1 - \frac{s}{\rho}\right) e^{s/\rho}. \tag{8.5.10}$$

The symmetry $\xi(s) = \xi(1-s)$ implies that roots $\neq 1/2$ come in pairs $\{\rho, 1-\rho\}$. Note that $|1-\rho| \sim |\rho|$ as $|\rho| \to \infty$.

If we pair the corresponding two factors in (8.5.10), we obtain

$$\begin{aligned}\left(1 - \frac{s}{\rho}\right) e^{s/\rho} \left(1 - \frac{s}{1-\rho}\right) e^{s/(1-\rho)} &= \frac{\rho - s}{\rho} \frac{1-\rho-s}{1-\rho} \exp\left(\frac{s}{\rho} + \frac{s}{1-\rho}\right) \\ &= \left(1 - \frac{s(1-s)}{\rho(1-\rho)}\right) \exp\left(\frac{s}{\rho(1-\rho)}\right).\end{aligned}$$

Now $1/\rho(1-\rho) = O(|\rho|^{-2})$. It follows that, paired this way, we may separate out the exponential terms and are left with absolutely convergent product

$$\exp\left(\sum \frac{s}{\rho(1-\rho)}\right) \prod \left(1 - \frac{s(1-s)}{\rho(1-\rho)}\right).$$

It follows from this and the factorization theorem that

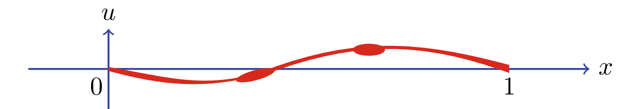



**Fig. 8.1** Inhomogeneous string

$$\xi(s) \;=\; e^{Q(s)} \prod_{\rho} \left(1 - \frac{s}{\rho}\right),$$

where the product is interpreted in the sense above, and $Q$ is a polynomial of degree at most 1. Both $\xi$ and the product are invariant under $s \to (1-s)$, so the same must be true of $Q$. Therefore $Q$ has degree zero, and we have (8.5.8). □

## 8.6 Application: an inhomogeneous vibrating string

Suppose that the density of a string is described by a function $m : [0,1] \to \rho_+$. We assume that there are constants $M_0$ and $M_1$ such that

$$0 < M_0 \;\leq\; m(x) \;\leq\; M_1 < \infty. \tag{8.6.1}$$

Let $u(x,t)$, $0 \leq x \leq 1$, $t \in \mathbb{R}$ denote the height of the (center of the cross-section of the) string above its resting position at height 0, with fixed endpoints $u(0,t) = u(1,t) = 0$; see Figure 8.1. The equation of motion of the vibrating string is given by Newton's second law: mass times acceleration equals force. The acceleration at time $t$, at the point $x$ is $u_{tt}(x,t)$ and the force, due to the tension, is proportional to $u_{xx}(x,t)$. Therefore (with a convenient choice of units), the string equation is

$$m(x)\, u_{tt}(x,t) \;=\; u_{xx}(x,t), \tag{8.6.2}$$

where $m$ is the mass density at $x$. We make no particular regularity assumptions on $m$—it could be piecewise continuous, or merely measurable.

A standard trick for such equations is to separate variables. One looks for solutions that have the form

$$u(x,t) \;=\; \varphi(x)\,\psi(t).$$

Computing (8.6.2) for this case and dividing by the product $m\varphi\psi$ give

$$\frac{\psi''}{\psi} \;=\; \frac{\varphi''}{m\varphi}.$$

By assumption, the left side is independent of $x$ and the right side is independent of $t$, so each quotient is equal to some constant $\lambda$. The resulting equation for $\psi$ puts no constraint on the constant $\lambda$, but the equation

$$\varphi''(x) = \lambda m(x)\,\varphi(x) \tag{8.6.3}$$

with boundary conditions

$$\varphi(0) = \varphi(1) = 0 \tag{8.6.4}$$

does constrain $\lambda$. Clearly $\lambda = 0$ corresponds to the trivial solution $\varphi \equiv 0$. As we shall see, there is a sequence of solutions $\{\varphi_n\}$ with

$$0 > \lambda_1 > \lambda_2 > \lambda_3 > \dots, \qquad \lambda_n \to -\infty.$$

Corresponding real-valued solutions of the (8.6.2) are then

$$u_n(x,t;a) = \cos(\sqrt{|\lambda_n|}\,t - a)\,\varphi_n(x), \qquad 0 \le a < 2\pi.$$

The first step is to make precise the meaning of the equation (8.6.3). If we assume for the moment that $\varphi$ has a continuous first derivative and $\varphi(0) = a$, $\varphi'(0) = b$, then the equation (8.6.3) can be taken to mean

$$\begin{aligned}
\varphi(x) &= a + \int_0^x \varphi'(t)\,dt \\
&= a + \int_0^x \left[ b + \int_0^t \varphi''(s)\,ds \right] dt \\
&= a + bx + \lambda \int_0^x \int_0^t m(s)\varphi(s)\,ds\,dt.
\end{aligned} \tag{8.6.5}$$

We define a solution of (8.6.3) to be a continuous function $\varphi : [0,1] \to \mathbb{C}$ that is a solution of (8.6.5) for some $a, b \in \mathbb{C}$.

**Proposition 8.6.1.** *For each pair $a, b \in \mathbb{C}$, there is a unique solution of the integral equation (8.6.5).*

*Proof:* The difference of two solutions of (8.6.5) with the same values $a, b$ is a solution with values $a = b = 0$. To prove uniqueness, therefore, it is enough to show that if $\varphi$ is a continuous solution with $a = b = 0$, then $\varphi \equiv 0$. Let $C$ be an upper bound for $|\varphi(x)|$, $0 \le x \le 1$. Then (8.6.5) and (8.6.1) imply

$$|\varphi(x)| \le C|\lambda M_1|\frac{x^2}{2}, \quad 0 \le x \le 1.$$

By induction

$$|\varphi(x)| \le \frac{C|\lambda M_1|^n x^{2n}}{(2n)!}, \qquad n = 1, 2, 3, \dots . \tag{8.6.6}$$

The right side has limit zero as $n \to \infty$.

Existence is proved by setting

$$\varphi_0(x) = a + bx, \qquad \varphi_{n+1}(x) = \int_0^x \int_0^t m(s)\varphi_n(s)\,ds\,dt.$$

Formally, the sum

$$\varphi(x,\lambda) \equiv a+bx+\sum_{n=1}^{\infty}\lambda^n\varphi_n(x) \tag{8.6.7}$$

is a solution of (8.6.5). Convergence of the series (8.6.7) follows from the estimates

$$|\varphi_n(x)| \le |a|\cdot\frac{M_1^n x^{2n}}{(2n)!}+|b|\cdot\frac{M_1^n x^{2n+1}}{(2n+1)!}; \tag{8.6.8}$$

see Exercises 11 and 12 for details. □

**Proposition 8.6.2.** *The values of $\lambda$ such that the problem (8.6.3), (8.6.5) has a non-zero solution form an infinite sequence $\{\lambda_n\}$ with*

$$0 > \lambda_1 > \lambda_2 > \lambda_3 > \dots, \qquad \lambda_n \to -\infty, \tag{8.6.9}$$

*such that*

$$\sum_{n=0}^{\infty}\frac{1}{|\lambda_n|^{1/2+\varepsilon}} < \infty, \qquad \text{all } \varepsilon>0. \tag{8.6.10}$$

*Proof:* If $\varphi$ is a non-zero solution, then it is a solution of (8.6.5) with $a=0$, $b\neq 0$, with the additional property that $\varphi(1)=0$. It can be normalized by taking $b=1$. Moreover, an integration by parts shows that

$$\lambda\int_0^1 m(x)|\varphi(x)|^2\,dx = \int_0^1\varphi''(x)\overline{\varphi(x)}\,dx = -\int_0^1|\varphi'(x)|^2\,dx,$$

so $\lambda<0$.

Let $\Phi(x,\lambda)$ be the solution of (8.6.5) with $a=0$, $b=1$. Then the normalized solutions of (8.6.3) and (8.6.4) are precisely the $\Phi(x,\lambda)$ for which $\Phi(1,\lambda)=0$. According to the estimates (8.6.6) and (8.6.8),

$$|\Phi(1,\lambda)| \le \frac{\sinh(|\lambda M_1|^{1/2})}{|\lambda M_1|^{1/2}} = O(e^{|\lambda M_1|^{1/2}}).$$

Thus $\Phi(1,\lambda)$ is an entire function of $\lambda$ of order at most $1/2$. On the other hand, it is easily established that for $\lambda>0$, the summands $\varphi_n$ in (8.6.7) satisfy

$$\varphi_n(x) \ge \frac{M_0^n x^{2n+1}}{(2n+1)!}. \tag{8.6.11}$$

It follows that for $\lambda>0$,

$$\Phi(1,\lambda) \ge \frac{\sinh(\lambda M_0)^{1/2})}{(\lambda M_0)^{1/2}}. \tag{8.6.12}$$

Therefore $\Phi(1,\lambda)$ is precisely of order $1/2$.

Note that $\Phi(1,0)=1$. By Theorem 8.4.1,

$$\Phi(1,\lambda) = \prod_{n=1}^{\infty}\left(1-\frac{\lambda}{\lambda_n}\right).$$

But $\Phi(1,\lambda)$ is not a polynomial, so the number of zeros $\{\lambda_n\}$ is infinite. Our argument so far shows that the zeros are negative. The uniqueness argument shows that an eigenfunction $\varphi_\nu$ is determined uniquely by $\varphi'_\nu(0)$, so any other eigenfunction for $\lambda_\nu$ is a constant multiple of $\varphi_\nu$. Thus the eigenvalues are simple. This establishes (8.6.9).

The estimate (8.6.10) follows from Proposition 8.3.1. □

Proposition 8.6.2 can be improved with some input from differential equations technique. Note that in the constant density case $m(x) \equiv m$ the solutions are

$$\varphi_n(x) = \sin(|\lambda_n|^{1/2}\pi x), \qquad \lambda_n = -\frac{(n\pi)^2}{m}. \tag{8.6.13}$$

This suggests the following.

**Proposition 8.6.3.** *The values $\lambda_n$ of (8.6.9) satisfy the estimates*

$$-\frac{(n\pi)^2}{M_0} \le \lambda_n \le -\frac{(n\pi)^2}{M_1}. \tag{8.6.14}$$

For the proof, see Exercise 13.

## Exercises

1. Show that any choice of $p_n$ such that $p_n/\log n \ge a > 0$ is sufficient to force convergence of the product in (8.1.2).
2. Show that the choice of $p_n$ in Exercise 1 is best possible, in the sense that if $p_n/\log n \to 0$ as $n \to \infty$, then there is a sequence $\{z_n\}$, such that $\{|z_n|\}$ is nondecreasing, $|z_n| \to \infty$, but the product in (8.1.2) does not converge.
3. Use Proposition 1.8.1 to prove (8.2.5).
4. Suppose that $f$ is a polynomial of degree $n$ and $f(0) \ne 0$. Show that Theorem 8.2.1 implies that $f$ has exactly $n$ zeros, repeated according to multiplicity. (Use the asymptotics of $f(z)$ as $z \to \infty$.)
5. Prove that an entire function of fractional order attains every finite value infinitely many times.
6. (a) Apply Hadamard's theorem to $\sin z$.

   (b) Apply Hadamard's theorem to $\cos z$.
7. Prove that the estimate (10.5.8) implies (8.5.7).
8. Prove that the zeros of $\xi$ satisfy

$$\sum \frac{1}{\rho \log \rho} < \infty.$$

9. (a) Show that the order of

$$f(z) = \sum_{n=0}^{\infty} \frac{z^n}{(n\,!)^{\alpha}}, \quad \alpha > 0$$

is at least $1/\alpha$. (Estimate, for fixed $r$, the maximum value of $r^n/(n\,!)^{\alpha}$.)

(b) Show that the order is $\leq 1/\alpha$. Use the estimate

$$f(r) < \sum_{n=N}^{\infty} \frac{z^n}{(n\,!)^{\alpha}} + \sum_{n=N+1}^{\infty} \frac{z^n}{[(N+1)\,!\,N^{n-N-1}]^{\alpha}}$$
$$< Cr^n + \frac{r^{N+1}}{[(N+1)\,!]^{\alpha}(1-r/N^{\alpha})}, \quad N^{\alpha} > r.$$

10. Show that $f(z) = \int_0^{\infty} e^{-t^2} \cos zt\, dt$ has order 2.
11. Prove the estimates (8.6.6), (8.6.8), and (8.6.11).
12. Verify that the series (8.6.7) is a solution of (8.6.5).
13. Suppose that $\varphi_1$ and $\varphi_2$ are non-zero solutions of two versions of (8.6.3):

$$\varphi_1'' = -\mu_1 \varphi_1, \qquad \varphi_2'' = -\mu_2 \varphi_2,$$

and suppose $\mu_2(x) > \mu_1(x) > 0$ for $0 \leq x \leq 1$.

(a) Show that the Wronskian $W = \varphi_1 \varphi_2' - \varphi_1' \varphi_2$ is either identically zero or is never zero. (Hint: look at $W'$.)

(b) Suppose that $\varphi_1(a) = \varphi_1(b) = 0$ for some points $a, b \in [0,1]$, $a < b$. Suppose also that $W \neq 0$. Show that if $\varphi_1$ has no zeros between $a$ and $b$, then $\varphi_2$ has a zero between $a$ and $b$. (Hint: $\varphi'(b) \neq 0$.) (This is a case of the Sturm Comparison Theorem.)

(c) Show that in the case of (8.6.3), $\lambda_n$ is the value of $\lambda$ for which $\Phi(x, \lambda)$ has exactly $n$ zeros in $(0,1]$, one of them at $x = 1$.

(d) Prove Proposition 8.6.3.

## Remarks and further reading

Entire functions are often considered in the context of entire meromorphic functions—see Chapter 9 and the references there. See also Boas [24] and Levin [88], [89].

# Chapter 9
# Value distribution theory



The first general result about values of a complex function is due to Gauss: a complex polynomial of degree $n$ takes every complex value exactly $n$ times, counting multiplicity. Early in the development of complex function theory it was known that each rational function, as a function on the Riemann sphere, takes every value (finite or infinite) the same number of times. A theorem of Picard says that a non-constant entire meromorphic function, i.e. a function meromorphic on all of $\mathbb{C}$, can omit at most two values. A related theorem says that such a function can have at most four values at which all the solutions are multiple—a result that severely limits the curves that can be parametrized by meromorphic functions.

In the 1920s Nevanlinna undertook a deep study of the distribution of values of entire meromorphic functions. Each of the results mentioned above, as well as a stronger version of Picard's theorem, can be proved using Nevanlinna's theory. This theory also leads to versions for entire meromorphic functions of Hadamard's theory of entire functions; see Chapter 8.

In this chapter we introduce the "Nevanlinna characteristic" of an entire meromorphic function, and a closely related version due (separately) to Ahlfors and to Shimizu. The first and second "fundamental theorems" are proved, and are then applied to obtain the results mentioned above. This is only a brief introduction to a large body of work, much of it of current interest; see the remarks at the end of the chapter.

## 9.1 The Nevanlinna characteristic and the first fundamental theorem

Throughout this section, $f$ denotes a non-constant entire meromorphic function. The integral

$$\frac{1}{2\pi}\int_0^{2\pi} \log|f(re^{i\theta})|\,d\theta \tag{9.1.1}$$



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_9

plays a central role. To analyze it we start by noting that the integrand is the real part of

$$\log f(re^{i\theta}) \;=\; \log|f(re^{i\theta})| + i\arg f(re^{i\theta}) \;=\; U(x,y) + iV(x,y).$$

Since

$$r\frac{\partial}{\partial r} \;=\; x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}, \qquad \frac{\partial}{\partial \theta} \;=\; -y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y},$$

the Cauchy–Riemann equations imply that

$$r\frac{\partial U}{\partial r} \;=\; xU_x + yU_y \;=\; xV_y - yV_x \;=\; \frac{\partial V}{\partial \theta}.$$

Therefore, assuming that $f$ has no zeros or poles on the circle $\{z : |z| = r\}$,

$$r\frac{d}{dr}\left\{\frac{1}{2\pi}\int_0^{2\pi} \log|f(re^{i\theta})|\,d\theta\right\} \;=\; \frac{1}{2\pi}\int_0^{2\pi} \frac{\partial}{\partial \theta}\left\{\arg f(re^{i\theta})\right\}d\theta. \tag{9.1.2}$$

The integral on the right is the number of zeros of $f$ in the disk $\{z : |z| < r\}$ minus the number of poles in the disk; see Exercise 1. In general, set

$$n(r,a,f) \;=\; \text{number of solutions of } f(z) = a, \quad |z| \le r, \tag{9.1.3}$$

counting multiplicity. Assume that $f(0) \neq 0,\infty$ and that $f(z) \neq 0,\infty$ when $|z| = r$. Then we have

$$\frac{1}{2\pi}\int_0^{2\pi} \log|f(re^{i\theta})|\,d\theta \;=\; \log|f(0)| + N(r,0,f) - N(r,\infty,f), \tag{9.1.4}$$

where

$$N(r,0,f) \;=\; \int_0^r \frac{n(s,0,f)}{s}\,ds, \qquad N(r,\infty,f) \;=\; \int_0^r \frac{n(s,\infty,f)}{s}\,ds. \tag{9.1.5}$$

Let us push this one step further by allowing for a zero or pole at $z = 0$. In each case $f$ has a Laurent expansion at $z = 0$,

$$f(z) \;=\; cz^p + \sum_{n>p} c_n z^n \;=\; z^p g(z).$$

We can apply the preceding argument to the entire meromorphic function $g$, leading to

$$\begin{aligned}\frac{1}{2\pi}\int_0^{2\pi} \log|f(re^{i\theta})|\,d\theta &= p\log r + \frac{1}{2\pi}\int_0^{2\pi} \log|g(re^{i\theta})|\,d\theta \\ &= p\log r + \log|c| + N(r,0,g) - N(r,\infty,g).\end{aligned} \tag{9.1.6}$$

Note that if $p \ge 0$ then $n(r,0,g) \;=\; n(r,0,f) - p$, while if $p < 0$, $n(r,\infty,g) = n(r,\infty,f) + p$. Thus in either case (9.1.6) is just (9.1.4) with $f(0)$ replaced by the leading coefficient in the Laurent expansion and (9.1.5) replaced by

$$N(r,a,f) = \int_0^r \frac{n(s,a,f)-n(0,a,f)}{s}\,ds + n(0,a,f)\log r. \tag{9.1.7}$$

At this point Nevanlinna's idea was to regroup the positive and negative terms. For positive $x$ let

$$\log^+ x = \max\{\log x, 0\}, \qquad \log^-(x) = \max\{-\log x, 0\} = \log^+(1/x).$$

Then $\log x = \log^+ x - \log^+(1/x)$. Let

$$m(r,a,f) = \frac{1}{2\pi}\int_0^{2\pi} \log^+ \frac{1}{|f(re^{i\theta})-a|}\,d\theta = m(r,0,f-a) \tag{9.1.8}$$

and

$$m(r,\infty,f) = \frac{1}{2\pi}\int_0^{2\pi} \log^+ |f(re^{i\theta})|\,d\theta. \tag{9.1.9}$$

Then (9.1.6) can be rewritten in the form

$$\log|c| + m(r,0,f) + N(r,0,f) = m(r,\infty,f) + N(r,\infty,f). \tag{9.1.10}$$

The term $m(r,0,f)$ is a measure of how often $|f(re^{i\theta})| < 1$, i.e. how often, on average, $f(re^{i\theta})$ is close to zero. The term $m(r,\infty,f)$ is a measure of how often, on average, $f(re^{i\theta})$ is close to $\infty$.

Clearly for each $a \in \mathbb{C}$,

$$n(r,a,f) = n(r,0,f-a), \qquad n(r,\infty,f) = n(r,\infty,f-a).$$

Therefore (9.1.10) can be put in a more general form.

**Proposition 9.1.1.** *Suppose that $f$ is an entire meromorphic function, and that $c(a)$ is the leading coefficient of the Laurent expansion of $f(z)-a$ at $z=0$. Then*

$$\log|c(a)| + T(r,a,f) = T(r,f-a), \tag{9.1.11}$$

*where*

$$\begin{aligned} T(r,a,f) &= m(r,a,f) + N(r,a,f), \\ T(r,f) &= m(r,\infty,f) + N(r,\infty,f). \end{aligned} \tag{9.1.12}$$

The function $T(r,a,f)$ is the *Nevanlinna characteristic* of $f$. Nevanlinna's "first fundamental theorem" says that $T(r,a,f) - T(r,f)$ is a bounded function of $r$. This will be proved in the next section.

For applications, it can be useful to pass to an equivalent formulation of (9.1.7) by integrating by parts and representing $N(r,a,f)$ as a Stieltjes integral (see Section 1.8):

$$
\begin{aligned}
N(r,a,f) &= \int_0^r \frac{n(s,a,f)-n(0,a,f)}{s}\,ds + n(0,a,f)\log r \\
&= [n(r,a,f)-n(0,a,f)]\log r - \int_0^r \log s\, dn(s,a,f) + n(0,a,f)\log r \\
&= n(r,a,f)\log r - \sum_{r_j<r} \log r_j. \qquad (9.1.13)
\end{aligned}
$$

Here $r_j = |z_j|$ and the $z_j$ are the solutions of $f(z_j) = a$, repeated according to multiplicity and numbered with $|z_j| \le |z_{j+1}|$.

The asymptotics of $T(r,a,f)$ as $r \to \infty$ are important.

**Proposition 9.1.2.** *Suppose that $f$ is a non-constant rational function, $f = P/Q$ where $P$ and $Q$ are polynomials with no common zeros. Then for each $a \in \mathbb{S}$,*

$$
\lim_{r\to\infty} \frac{T(r,a,f)}{\log r} = \max\{\deg P, \deg Q\}. \qquad (9.1.14)
$$

This follows easily from the definition of $T$, and (9.1.13); see Exercise 6.

**Proposition 9.1.3.** *The limit*

$$
\lim_{r\to\infty} \frac{T(r,f)}{\log r} \qquad (9.1.15)
$$

*is finite if and only if $f$ is rational.*

*Proof:* The necessity of the condition follows from Proposition 9.1.2. If the limit (9.1.15) is finite, then each $N(r,\infty,f)$ is finite: $f$ has only finitely many poles. Therefore $z = \infty$ is an isolated singularity of $f$. Suppose that $f$ is not rational. Then $\infty$ is an essential singularity. By the Casorati–Weierstrass theorem, the image of each set $\Omega_n = \{z : n < |z| < \infty\}$ is dense in $\mathbb{C}$. It follows that we may choose a decreasing sequence of closed disks $A_n \subset f(\Omega_n)$. The centers of these disks converge to a finite point $a$ that belongs to each $A_n$. Then $N(r,a,f)/\log r$ has limit $\infty$, so the limit (9.1.15) is also infinite. □

It follows from (9.1.13) that an equivalent form of (9.1.11) is

$$
\begin{aligned}
&\log|c(a)| + m(r,a,f) + n(r,a,f)\log r - \sum_{r_j<r} \log r_j \\
&\quad = m(r,\infty,f-a) + n(r,\infty,f)\log r - \sum_{s_j<r} \log s_j, \qquad (9.1.16)
\end{aligned}
$$

where $s_j = |w_j|$ and the $w_j$ are the poles of $f$, repeated according to multiplicity. Note that

$$
-m(r,a,f) + m(r,\infty,f-a) = \frac{1}{2\pi}\int_0^{2\pi} \log|f(re^{i\theta}) - a|\,d\theta.
$$

Therefore when $a = 0$, $f(0) \neq 0$ and $f$ has no poles in the closed disk $\{z : |z| \leq r\}$, (9.1.16) reduces to

$$\log|f(0)| + n(r,0,f)\log r - \sum_{r_j \leq r} \log r_j \;=\; \frac{1}{2\pi}\int_0^{2\pi} \log|f(re^{i\theta})|\,d\theta,$$

which is Jensen's theorem, Theorem 8.2.1. Thus (9.1.16) is simply Jensen's theorem generalized to the situation when $f$ is meromorphic and $f-a$ may have a zero or pole at $z = 0$.

The first several results mentioned in the introduction can be deduced from (9.1.16); see the exercises.

## 9.2 The first fundamental theorem and a modified characteristic

Nevanlinna [105] proved two "fundamental theorems" about the characteristic $T$ that have striking applications to the study of the values of meromorphic functions. In this section we prove the first fundamental theorem, as well as a more elegant form that uses a modified version of the characteristic.

**Theorem 9.2.1.** (First fundamental theorem) *The difference $|T(r,a,f) - T(r,f)|$ is a bounded function of $r$:*

$$|T(r,a,f) - T(r,f)| \;\leq\; |\log|c(a)|| + \log^+|a| + \log 2. \tag{9.2.1}$$

*Proof:* It follows from (9.1.11) that

$$\begin{aligned} T(r,a,f) - T(r,f) \;&=\; -\log|c(a)| + |m(r,\infty,f-a) - m(r,\infty,f)| \\ &=\; -\log|c(a)| + \frac{1}{2\pi}\int_0^{2\pi}\Big[\log^+|f(re^{i\theta}) - a| - \log^+|f(re^{i\theta})|\Big]\,d\theta. \end{aligned} \tag{9.2.2}$$

Now for $x, y \geq 0$,

$$\log^+(x+y) \;\leq\; \log^+ x + \log^+ y + \log 2;$$

this is a special case of Exercise 7. Moreover $\log^+ x$ is non-decreasing for $x \geq 0$. Therefore

$$\log^+|f(z) - a| \;\leq\; \log^+(|f(z)| + |a|) \;\leq\; \log^+|f(z)| + \log^+|a| + \log 2.$$

The same argument, with $f$ and $f-a$ interchanged, shows that the integrand in (9.2.2) is dominated by $\log^+|a| + \log 2$. Integrating gives (9.2.1). □

**Corollary 9.2.2.** *For every $a \in \mathbb{S}$,*

$$\lim_{r\to\infty} \frac{T(r,a,f)}{\log r} \;\geq\; 1.$$

*Proof:* This follows easily from Theorem 9.2.1 and Proposition 9.1.2 for rational $f$, or Proposition 9.1.3 for irrational $f$. □

It will be convenient to work with a modified version of the characteristic, due to Ahlfors and to Shimizu. This version takes advantage of the distance function that is obtained by identifying the plane with the unit sphere (minus the north pole) via the inverse of the stereographic projection; see Section 2.1:

$$d(z,w) = \frac{|z-w|}{\sqrt{(1+|z|^2)(1+|w|^2)}}. \tag{9.2.3}$$

In particular

$$d(z,0) = \frac{|z|}{\sqrt{1+|z|^2}}, \qquad d(z,\infty) = \frac{1}{\sqrt{1+|z|^2}}. \tag{9.2.4}$$

To obtain a spherical version of the function $m$, defined by (9.1.8), we replace $|f(re^{i\theta})-a|$ by the spherical distance (9.2.3) and replace the cutoff for $|f(re^{i\theta})-a|<1$ by a comparison of the distance $d(f(re^{i\theta}),a)$ to the distance $d(f(0),a)$:

$$m^\circ(r,a,f) = \frac{1}{2\pi}\int_0^{2\pi} \log\frac{d(f(0),a)}{d(f(re^{i\theta}),a)}\,d\theta, \quad \text{if } f(0)\neq a. \tag{9.2.5}$$

Then

$$\lim_{r\to 0} m^\circ(r,a,f) = 0 \quad \text{if} \quad f(0)\neq a. \tag{9.2.6}$$

Modifications are needed to cover the cases $f(0)=a\in\mathbb{C}$ and $f(0)=\infty$. First,

$$m^\circ(r,a,f) = \frac{1}{2\pi}\int_0^{2\pi} \log\frac{|c|}{(1+|a|^2)d(f(re^{i\theta}),a)}\,d\theta, \quad \text{if} \quad f(0)=a\in\mathbb{C}, \tag{9.2.7}$$

where $c$ is the leading coefficient in the Taylor expansion of $f(z)-a$ at $z=0$.

Second,

$$m^\circ(r,\infty,f) = \frac{1}{2\pi}\int_0^{2\pi} \log\frac{\sqrt{1+|f(re^{i\theta})|^2}}{|c|}\,d\theta, \quad \text{if } f(0)=\infty, \tag{9.2.8}$$

where $c$ is the leading coefficient in the Laurent expansion of $f$ at $z=0$.

The corresponding modification of the Nevanlinna characteristic is the *Ahlfors–Shimizu characteristic*

$$T^\circ(r,a,f) = m^\circ(r,a,f)+N(r,a,f), \tag{9.2.9}$$

where $N(r,a,f)$ is defined, as before, by (9.1.7). For $T^\circ$ the first fundamental theorem takes an elegant form.

**Theorem 9.2.3.** *The Ahlfors–Shimizu characteristic $T^\circ(r,a,f)$ is independent of $a\in\mathbb{S}=\mathbb{C}\cup\{\infty\}$.*

*Proof:* Suppose first that $a \neq b$ and $f(0) \neq a,b$. Then

$$m^\circ(r,a,f) - m^\circ(r,b,f) \;=\; \frac{1}{2\pi}\int_0^{2\pi} \log\left[\frac{d(f(0),a)\,|f(re^{i\theta})-b|}{d(f(0),b)\,|f(re^{i\theta})-a|}\frac{\sqrt{1+a^2}}{\sqrt{1+b^2}}\right] d\theta.$$

As in the computation leading to (9.1.2) and (9.1.4),

$$\begin{aligned} r\frac{d}{dr}\{m^\circ(r,a,f)-m^\circ(r,b,f)\} &= \frac{1}{2\pi}\int_0^{2\pi} \frac{\partial}{\partial\theta}\left\{\arg\frac{f(re^{i\theta})-b}{f(re^{i\theta})-a}\right\} d\theta \\ &= n(r,b,f)-n(r,a,f). \end{aligned}$$

In view of (9.2.6), dividing by $r$ and integrating give

$$m^\circ(r,a,f) - m^\circ(r,b,f) \;=\; N(r,b,f) - N(r,a,f),$$

so

$$T^\circ(r,a,f) = T^\circ(r,b,f).$$

The remaining cases $f(0) = a \in \mathbb{C}$ or $f(0) = a = \infty$ are handled in the same way, using the modifications (9.2.7) or (9.2.8), respectively. □

In view of Theorem 9.2.3, we may write simply $T^\circ(r,f)$ for the modified characteristic.

**Theorem 9.2.4.** *The Nevanlinna characteristic $T(r,a,f)$ and the Ahlfors–Shimizu characteristic $T^\circ(r,f)$ differ by a bounded function of $r$.*

*Proof:* Note that

$$m(r,a,f) \;=\; m\left(r,\infty,\frac{1}{f-a}\right), \qquad m^\circ(r,a,f) \;=\; m^\circ\left(r,\infty,\frac{1}{f-a}\right).$$

The first of these identities follows immediately from the definitions, and the second follows from the fact that

$$d(a,0) \;=\; d\left(\frac{1}{a},\infty\right).$$

Therefore it is enough to examine $m^\circ(r,\infty,f) - m(r,\infty,f)$. Now

$$\frac{d(f(0),\infty)}{d(f(re^{i\theta}),\infty)} \;=\; \frac{\sqrt{1+|f(re^{i\theta})|^2}}{\sqrt{1+|f(0)|^2}}.$$

Let us split the interval $I = (0,2\pi)$ into the part $I_-$ where $|f(re^{i\theta})| \leq 1$ and its complement $I_+$. On $I_-$, $\log^+|f(re^{i\theta})| = 0$ and

$$\log\frac{1+|f(re^{i\theta})|^2}{\sqrt{1+|f(0)|^2}} \;\leq\; \log 2 - \log\sqrt{1+|f(0)|^2}.$$

On $I_+$, $\sqrt{1+|f(re^{i\theta})|^2} \leq 2|f(re^{i\theta})|$, so

$$\log^+ \frac{|f(re^{i\theta})|}{\sqrt{1+|f(0)|^2}} \leq \log \frac{\sqrt{1+|f(re^{i\theta})|^2}}{\sqrt{1+|f(0)|^2}}$$
$$\leq \log^+ |f(re^{i\theta})| + \log 2 - \log \sqrt{1+|f(0)|^2}.$$

Therefore

$$|m^\circ(r,\infty,f) - m(r,\infty,f)| \;\leq\; \left|\log 2 - \log \sqrt{1+|f(0)|^2}\right|. \qquad \square \tag{9.2.10}$$

Later we shall use an important consequence of the inequality (9.2.10). Although $m^\circ(r,a,f)$ may be negative, it follows from Theorem 9.2.4, and the fact that $m(r,a,f)$ is non-negative by definition, that

$$m^\circ(r,a,f) + O(1) \;\geq\; 0 \qquad \text{as } r \to \infty. \tag{9.2.11}$$

Theorem 9.2.4 allows us to carry Corollary 9.2.2 over to the modified characteristic:

$$\lim_{r\to\infty} \frac{T^\circ(r,f)}{\log r} \;\geq\; 1. \tag{9.2.12}$$

## 9.3 The second fundamental theorem

The second fundamental theorem provides an estimate for the sum of several integral terms $m(r,a_j,f)$ in terms of the single characteristic $T(r,f)$. This leads to deep results, such as Picard's theorem. Several proofs are known of Nevanlinna's original formulation. They have been described as either elementary, but far from simple, or simple, but far from elementary. Here we present the proof due to Ahlfors [4] for the Ahlfors–Shimizu characteristic $T^\circ(r,f)$, which lies between these extremes.

**Proposition 9.3.1.** *Suppose that $f$ is a non-constant entire meromorphic function. Then $T^\circ$ is strictly increasing with $r$ and*

$$\liminf_{r\to\infty} \frac{T^\circ(r,f)}{\log r} \;\geq\; 1. \tag{9.3.1}$$

*Proof:* Let us average (9.2.9) with respect to the planar measure $dm$ that corresponds to surface measure on the sphere under stereographic projection (see Exercise 8) of Chapter 2:

$$dm(z) \;=\; \frac{dx\,dy}{(1+|z|^2)^2}.$$

Averaging any function of $d(z,a)$ with respect to $a \in \mathbb{C}$ gives a result that is independent of $z$, so

$$\int_{\mathbb{C}} m^\circ(r,a,f)\,dm(a) \;=\; 0.$$

Therefore

$$T^\circ(r,f) \;=\; \frac{1}{\pi}\int_{\mathbb{C}} N(r,a,f)\,dm(a) \;\geq\; 0. \tag{9.3.2}$$

For non-constant $f$, the average of $n(r,a,f)$ with respect to $a$ is clearly strictly increasing with $r$, so the same is true of the average of $N(r,a,f)$.

Let $a_0 = f(0)$. Since $f$ is not constant, there is a $\delta > 0$ such that $f$ takes on every value $a$ such that $|a-a_0| < \delta$. Therefore $n(r_0,a,f)$ is eventually positive for such $a$, and for $r > r_0$,

$$N(r,a,f) \;\geq\; \int_{r_0}^{r} \frac{n(s,a,f)}{s}\,ds \;\geq\; \log r - \log r_0. \tag{9.3.3}$$

Averaging and taking the limit give the inequality (9.3.1). □

Differentiating (9.3.2) gives

$$r\frac{d}{dr}[T^\circ(r,f)] \;=\; \frac{1}{\pi}\int_{\mathbb{C}} n(r,a,f)\,dm(a).$$

This integral has the following interpretation. The set $\{a : n(r,a,f) = k\}$ is a portion of the image under $f$ of the closed disk $\{z : |z| \leq r\}$. Multiply the integral over this portion of the disk by $k$ and sum over $k$ to see that the integral represents the total area $A(r,f)$ of the image of the disk, $f(\{z : |z| \leq r\})$. Therefore

$$T^\circ(r,f) \;=\; \int_0^r \frac{A(s,f)}{s}\,ds, \tag{9.3.4}$$

with

$$A(r,f) \;=\; \int_{|a|\leq r} \frac{|f'(a)|^2}{[1+|f(a)|^2]^2}\,dm(a). \tag{9.3.5}$$

Note that the integrand remains bounded at the poles of $f$ and, in fact, is continuous.

We shall want to take averages with respect to a non-negative weight function $\rho$ of total weight 1:

$$\int_{\mathbb{C}} \rho(a)\,dm(a) \;=\; 1.$$

We shall assume that $\rho$ is chosen to be positive and continuous, except at isolated singularities, at which it is integrable.

Averaging $T^\circ(r,f)$ with respect to $\rho$,

$$T^\circ(r,f) \;=\; \int_{\mathbb{C}} T^\circ(r,f)\rho(a)\,dm(a) \;=\; m_\rho(r,f) + N_\rho(r,f) + c(a,f), \tag{9.3.6}$$

where

$$N_\rho(r,f) \;=\; \int_{\mathbb{C}} N(r,a,f)\rho(a)\,dm(a) \tag{9.3.7}$$

and

$$\begin{aligned} m_\rho(r,f) &= \int_{\mathbb{C}} \left\{ \frac{1}{2\pi} \int_0^{2\pi} \log \frac{1}{d(f(re^{i\theta}),a)}\,d\theta \right\} \rho(a)\,dm(a) \\ &= \int_{\mathbb{C}} m^\circ(r,a,f)\rho(a)\,dm(a) - c(a,f), \end{aligned} \tag{9.3.8}$$

where

$$c(a,f) \;=\; \int_{\mathbb{C}} \log[d(f(0),a)]\,\rho(a)\,dm(a)$$

(since $f(0)$ is the leading coefficient in the Laurent expansion of $f(z)-a$ at zero, with the exception of one value of $a$).

Note that, given $w$, the (normalized) chord length $d(w,a)$ is less than 1 except for one value of $a$, so

$$m_\rho(r,f) \;>\; 0. \tag{9.3.9}$$

Let

$$n_\rho(r,f) \;=\; r\frac{d}{dr}[N_\rho(r,f)] \;=\; \int_{\mathbb{C}} n(r,a,f)\rho(a)\,dm(a).$$

The integral here is the integral of $f$ over the image $f(\{z : |z| \le r|\})$ as described above, so

$$n_\rho(r,f) \;=\; \int_{|a|<r} \frac{|f'(a)|^2}{[1+|f(a)|^2]^2}\,\rho(f(a))\,dm(a) \;=\; \int_0^r \lambda(s,f)s\,ds,$$

where

$$\lambda(r,f) \;=\; \int_0^{2\pi} \frac{|f'(re^{i\theta})|^2}{[1+|f(re^{i\theta})|^2]^2}\,\rho(f(re^{i\theta}))\,d\theta. \tag{9.3.10}$$

Because of the assumptions on $\rho$ and the remark after (9.3.5), $\lambda(r,f)$ is continuous and is positive for $r>0$. For any choice of $0<r_0<r$,

$$N_\rho(r,f) - N_\rho(r_0,f) \;=\; \int_{r_0}^r \left\{ \int_0^s \lambda(t,f)t\,dt \right\} \frac{ds}{s}. \tag{9.3.11}$$

Taking into account (9.3.6), Proposition 9.3.1, and (9.3.9), we have

$$\begin{aligned} T^\circ(r,f) \ge T^\circ(r,f) - T^\circ(r_0,f) \;&=\; m_\rho(r,f) - m_\rho(r_0,f) + N_\rho(r,f) - N_\rho(r_0,f) \\ &> \int_{r_0}^r \left\{ \int_0^s \lambda(t,f)t\,dt \right\} \frac{ds}{s} - m_\rho(r_0,f). \end{aligned} \tag{9.3.12}$$

The next step in the analysis is to examine $\lambda(r,f)$ more closely. The following well-known inequality will be used twice. (This is a special case of Jensen's inequality [73]; see Exercise 16.)

**Lemma 9.3.2.** *If $f$ is non-negative and integrable on the interval $[a,b]$, then*

$$\log\left\{\frac{1}{b-a}\int_a^b f(x)\,dx\right\} \;\geq\; \frac{1}{b-a}\int_a^b \log f(x)\,dx. \tag{9.3.13}$$

Applying Lemma 9.3.2 to $\lambda(r,f)/2\pi$ in (9.3.10) gives

$$\begin{aligned}\log\lambda(r,f)-\log 2\pi \;\geq\; & \frac{1}{2\pi}\int_0^{2\pi}\log\rho(f(re^{i\theta}))\,d\theta \\ & +\frac{1}{2\pi}\int_0^{2\pi}\log\frac{|f'(re^{i\theta})|^2}{[1+|f(re^{i\theta})|^2]^2}\,d\theta.\end{aligned} \tag{9.3.14}$$

Let

$$\mu(r,f) \;=\; \frac{1}{2\pi}\int_0^{2\pi}\log\frac{|f'(re^{i\theta})|}{1+|f(re^{i\theta})|^2}\,d\theta,$$

so that (9.3.14) is

$$\log\lambda(r,f) \;\geq\; \frac{1}{2\pi}\int_0^{2\pi}\log\rho(f(re^{i\theta}))\,d\theta+2\mu(r,f)+\log 2\pi. \tag{9.3.15}$$

Now

$$\mu(r,f) \;=\; \frac{1}{2\pi}\int_0^{2\pi}\log|f'(re^{i\theta})|\,d\theta+\frac{1}{2\pi}\int_0^{2\pi}\log\frac{1}{1+|f(re^{i\theta})|^2}\,d\theta. \tag{9.3.16}$$

Applying the argument used in the proof of (9.1.2),

$$r\frac{d}{dr}\left\{\frac{1}{2\pi}\int_0^{2\pi}\log|f'(re^{i\theta})|\,d\theta\right\} \;=\; n(r,0,f')-n(r,\infty,f'). \tag{9.3.17}$$

The integrand of the second term on the right in (9.3.16) is

$$\begin{aligned}& 2\log\frac{[1+|f(0)|^2]^{1/2}}{[1+|f(re^{i\theta})|^2]^{1/2}}-2\log\frac{1}{[1+|f(0)|^2]^{1/2}} \\ &\quad=\; -2\log\frac{d(f(0),\infty)}{d(f(re^{i\theta}),\infty)}-2\log\frac{1}{[1+|f(0)|^2]^{1/2}}.\end{aligned}$$

Therefore

$$\begin{aligned}\frac{d}{dr}[\mu(r,f)] \;&=\; \frac{1}{r}\left[n(r,0,f')-n(r,\infty,f')\right]-2\frac{d}{dr}[m^\circ(r,\infty,f)] \\ &= \frac{1}{r}\left[n(r,0,f')-n(r,\infty,f')\right]-2\frac{d}{dr}\left[T^\circ(r,f)-N(r,\infty,f)\right] \\ &= \frac{1}{r}\left[n(r,0,f')-n(r,\infty,f')+2n(r,\infty,f)\right]-2\frac{d}{dr}[T^\circ(r,f)].\end{aligned} \tag{9.3.18}$$

Let us examine

$$n_1(r,f) \equiv n(r,0,f') - n(r,\infty,f') + 2n(r,\infty,f). \tag{9.3.19}$$

The first summand is a sum over the multiple zeros $a$ of $f$, $|a| < r$, and assigns $k-1$ to each zero of multiplicity $k$. The second summand is a sum over poles $a$ of $f$, $|a| < r$, that assigns $-(k+1)$ to each pole of multiplicity $k$. The last summand assigns $2k$ to each such pole. Therefore the counting function $n_1$ assigns $k-1$ to each zero or pole $a$, $|a| < r$, having multiplicity $k$.

Integrating (9.3.18), we obtain (up to an additive constant)

$$\mu(r,f) = N_1(r,f) - 2T^\circ(r,f), \tag{9.3.20}$$

where

$$N_1(r,f) = N(r,0,f') - N(r,\infty,f') + 2N(r,\infty,f).$$

Combining (9.3.15) and (9.3.20), we obtain

$$\begin{aligned} &\frac{1}{2\pi}\int_0^{2\pi} \log\rho(f(re^{i\theta}))\,d\theta \\ &\leq 4T^\circ(r,f) - 2N_1(r,f) + \log\lambda(r,f) - \log 2\pi. \end{aligned} \tag{9.3.21}$$

The proof of the second fundamental theorem involves taking advantage of the inequality (9.3.21) by making a careful choice of the density function $\rho$. Given distinct points $a_1, a_2, \dots, a_q$ in $\mathbb{S}$, let

$$\log\rho(a) = 2\sum_{j=1}^{q} \log\frac{1}{d(a,a_j)} - 2\log\left\{\sum_{j=1}^{q}\log\frac{1}{d(a,a_j)}\right\} - 2C, \tag{9.3.22}$$

where $C$ is chosen so that $\int_{\mathbb{C}} \rho(a)\,dm(a) = 1$. Note that $\rho$ is integrable at the singularities $a_j$: as $t = d(a,a_j) \to 0$,

$$\rho(a) \sim \frac{1}{t^2(\log t)^2},$$

which is integrable. In fact using polar coordinates centered at $a_j$, this comes down to integrability in one variable of $1/r(\log r)^2$ at $r = 0$. But this is the derivative of $-1/\log r$.

With this choice of $\rho$, making use again of Lemma 9.3.2,

$$\begin{aligned} &\frac{1}{2\pi}\int_0^{2\pi} \log\rho(f(re^{i\theta}))\,d\theta \\ &= 2\sum_{j=1}^{q}\frac{1}{2\pi}\int_0^{2\pi}\log\frac{1}{d(f(re^{i\theta}),a_j)}\,d\theta \\ &\quad -2\frac{1}{2\pi}\int_0^{2\pi}\log\sum_{j=1}^{q}\log\frac{1}{d(f(re^{i\theta}),a_j)}\,d\theta - 2C \end{aligned}$$

$$= 2\sum_{j=1}^{q} m^\circ(r,a_j,f) - 2\frac{1}{2\pi}\int_0^{2\pi} \log\sum_{j=1}^{q}\log\frac{1}{d(f(re^{i\theta}),a_j)}\,d\theta - 2C_1$$

$$\geq 2\sum_{j=1}^{q} m^\circ(r,a_j,f) - 2\log\left\{\sum_{j=1}^{q}\frac{1}{2\pi}\int_0^{2\pi}\log\frac{1}{d(f(re^{i\theta}),a_j)}\,d\theta\right\} - 2C_1$$

$$= 2\sum_{j=1}^{q} m^\circ(r,a_j,f) - 2\log\left\{\sum_{j=1}^{q} m^\circ(r,a_j,f)\right\} - 2C_2$$

$$\geq 2\sum_{j=1}^{q} m^\circ(r,a_j,f) - 2\log T^\circ(r,f) + O(1).$$

Combining this estimate with (9.3.21), we have

$$\sum_{j=1}^{q} m^\circ(r,a_j,f) \leq 2T^\circ(r,f) - N_1(r,f) + \log T^\circ(r,f) + \frac{1}{2}\log\lambda(r,f) + O(1). \tag{9.3.23}$$

This inequality brings us close to the spherical version of Nevanlinna's second major result.

**Theorem 9.3.3.** (Second fundamental theorem) *Suppose that $f$ is an entire meromorphic function. For distinct points $a_1, a_2, \ldots a_q$ in $\mathbb{S}$, the inequality*

$$\sum_{j=1}^{q} m^\circ(r,a_j,f) \; < \; 2T^\circ(r,f) - N_1(r,f) + \frac{\log r}{2} + O[\log T^\circ(r,f)] \tag{9.3.24}$$

*holds for each $r$ outside an open set $\Delta$ such that $\int_\Delta dr < \infty$.*

*Proof:* In view of (9.3.23), we only need to find an appropriate bound for $\lambda(r,f)$. Let

$$L(r,f) \; = \; \int_0^r \lambda(s,f)s\,ds, \qquad K(r,f) \; = \; \int_{r_0}^r \frac{L(s,f)}{s}\,ds.$$

Then (9.3.12) is

$$K(r,f) \; \leq T^\circ(r,f) + c, \tag{9.3.25}$$

for some constant $c = c(f)$.

Let $\Delta_1$ be the set of $r$ for which

$$\lambda(r,f) \; > \; r^{-1}L(r,f)^2. \tag{9.3.26}$$

For such $r$, $1 < r\lambda/L^2 = (dL/dr)L^{-2} = -d(L^{-1})/dr$, so

$$\int_{\Delta_1} dr < -\int_{\Delta_1}\frac{d}{dr}\left\{\frac{1}{L(r,f)}\right\}dr \; < \; \frac{1}{L(r_1,f)}, \tag{9.3.27}$$

where $r_1$ is the greatest lower bound of $\Delta_1$.

Similarly, let $\Delta_2$ be the set of $r$ for which

$$L(r,f) > rK(r,f)^2. \tag{9.3.28}$$

For such $r$, $1 < L/rK^2 = (dK/dr)K^{-2} = -d(K^{-1})/dr$, so

$$\int_{\Delta_2} dr < -\int_{\Delta_2} \frac{d}{dr}\left\{\frac{1}{K(r,f)}\right\} dr < \frac{1}{K(r_2,f)}, \tag{9.3.29}$$

where $r_2$ is the greatest lower bound of $\Delta_2$.

For $r$ in the complement of $\Delta = \Delta_1 \cup \Delta_2$,

$$\lambda(r,f) \leq \frac{L(r,f)^2}{r} \leq \frac{[rK(r,f)^2]^2}{r} = rK(r,f)^4. \tag{9.3.30}$$

By (9.3.1) and (9.3.25), the inequality (9.3.30) implies that

$$\lambda(r,f) = O[rT^\circ(r,f)^4 + c_1],$$

so

$$\log\lambda(r,f) = \log r + O[\log T^\circ(r,f)]. \tag{9.3.31}$$

Combining (9.3.31) with (9.3.23), we obtain (9.3.24). □

Theorem 9.2.4, Theorem 9.2.1, and Theorem 9.3.3 imply a more standard form of the second fundamental theorem.

**Theorem 9.3.4.** *Suppose that $f$ is an entire meromorphic function. For distinct points $a_1, a_2, \ldots a_q$ in $\mathbb{S}$, the inequality*

$$\sum_{j=1}^{q} m(r,a_j,f) < 2T(r,f) - N_1(r,f) + \frac{\log r}{2} + O[\log T(r,f)] \tag{9.3.32}$$

*holds for each $r$ outside an open set $\Delta$ such that $\int_\Delta dr < \infty$.*

**Remark.** For a somewhat different version of the Second Fundamental Theorem, see Theorem 9.5.1.

## 9.4 Applications

The second fundamental theorem implies two deep results of Picard. The first is better known in the equivalent form: a non-constant entire function can omit at most one value; see Exercise 17. This result is sometimes called Picard's "little" theorem to distinguish it from the "big" version in Exercise 24 of Chapter 17. For a different proof of the "little" theorem, see Section 17.3.

**Theorem 9.4.1.** *A non-constant entire meromorphic function can omit at most two distinct values.*

*Proof:* Suppose that $f$ omits distinct values $a_1, a_2, \dots, a_q$. Then

$$m^\circ(r, a_j, f) \;=\; m^\circ(r, a_j, f) + N(r, a_j, f) \;=\; T^\circ(r, f).$$

By Theorem 9.3.3, for $r$ in the complement of a set $\Delta$ such that $\int_\Delta dr < \infty$,

$$q \;\leq\; 2 + \frac{\log r}{2\,T^\circ(r, f)} + O\left(\frac{\log T^\circ(r, f)}{T^\circ(r, f)}\right).$$

By (9.2.12) we may choose a sequence of such values $r_n \to \infty$ such that the limit of the expression on the right is at most $2 + 1/2$. □

Note that the argument proves a stronger result: that there are at most two values $a_j$ for which

$$\limsup_{r \to \infty} \frac{N_1(r, a_j, f)}{T^\circ(r, f)} \;=\; 0, \tag{9.4.1}$$

where $N_1(r, a, f)$ comes from the counting function $n_1(r, a, f)$ that assigns $k-1$ to each solution $z$ of $f(z)$, $|z| < r$ that has multiplicity $k$. In fact, compare the sum of such terms to $N_1(r, f)$ and use (9.3.24).

The second theorem of Picard that was mentioned above has to do with algebraic curves

$$w^2 \;=\; P(z), \tag{9.4.2}$$

where $P$ is a polynomial of degree at least two, with no multiple zeros. Under a change of variables by a linear fractional transformation, a polynomial of given degree can be reduced to a canonical form. The first three are usually taken to be

$$P(z) = 1 - z^2; \tag{9.4.3}$$

$$P(z) = 4z^3 - g_2 z - g_3, \quad g_j \neq 0; \tag{9.4.4}$$

$$P(z) = (1 - z^2)(1 - k^2 z^2), \quad k^2 \neq 1. \tag{9.4.5}$$

In case (9.4.3), the curve (9.4.2) may be parametrized by setting

$$w(t) \;=\; \cos t, \qquad z(t) \;=\; \sin t.$$

As shown in Chapters 16 and 15, respectively, the curve can be parametrized in case (9.4.4) by

$$w(t) \;=\; \wp'(t), \qquad z(t) \;=\; \wp(t),$$

where $\wp$ is the Weierstrass $\wp$ function, and can be parametrized in case (9.4.5) by the Jacobi function sn:

$$w(t) \;=\; \mathrm{sn}'(t), \qquad z(t) \;=\; \mathrm{sn}\,(t).$$

A common feature here is that each of the functions used in the parametrization is an entire meromorphic function.

The *genus* of a curve (9.4.2) is defined to be $g$ if the degree of the polynomial $P$ is $2g+1$ or $2g+2$, so the curves we have considered so far have genus 0 and 1. The curve (9.4.2) is said to be *hyperelliptic* if it has genus $g > 1$.

**Theorem 9.4.2.** (Picard) *A hyperelliptic curve cannot be parametrized by entire meromorphic functions.*

We begin with an auxiliary result. A value $a$ is said to be *completely ramified* with respect to $f$ if every root of $f(z) = a$ is a multiple root. The Jacobi function sn and the Weierstrass function $\wp$ are examples of entire meromorphic functions with four completely ramified values; see the discussion of (9.4.8).

**Theorem 9.4.3.** *Suppose that $f$ is an entire meromorphic function that is not rational. Then $f$ has at most four completely ramified values.*

*Proof:* Let $n_1(r,a,f)$ be the counting function introduced in the discussion of (9.3.20) that assigns $k-1$ to each solution $z$ of $f(z) = a$, $|z| < r$, that has multiplicity $k$. If $a$ is completely ramified with respect to $f$, then either $f(z) = a$ has no roots or each root is multiple. In either case it follows that

$$n_1(r,a,f) \;\geq\; \frac{n(r,a,f)}{2}.$$

Therefore the corresponding integral $N_1$ of $n_1(s)/s$ satisfies

$$N_1(r,a,f) \;\geq\; \frac{N(r,a,f)}{2}.$$

It follows from this and (9.2.11) that

$$\begin{aligned} m^\circ(r,a,f) + N_1(r,a,f) &\geq \frac{1}{2}\left[m^\circ(r,a,f) + N(r,a,f) + O(1)\right] \\ &= \frac{1}{2}T^\circ(r,f) + O(1). \end{aligned} \tag{9.4.6}$$

Summing the inequalities (9.3.24) and (9.4.6) over completely ramified values $a_1,\dots,a_q$, we obtain, for most values of $r$,

$$\begin{aligned} \frac{q}{2}T^\circ(r,f) &= \sum_{j=1}^{q}\left[m^\circ(r,a_j,f) + N_1(r,a_j,f)\right] + O(1) \\ &\leq 2T^\circ(r,f) + \frac{\log r}{2} + O(|\log T^\circ(r,f)|). \end{aligned} \tag{9.4.7}$$

Now if $f$ has any completely ramified values, then

$$\lim_{r\to\infty} \frac{T^\circ(r,f)}{\log r} \;\geq\; 2;$$

see Exercise 13. Therefore, dividing (9.4.7) by $T^\circ(r,f)$ and letting $r\to\infty$, we obtain $q \le 4+1/2$. □

*Proof of Theorem* 9.4.2. Suppose that

$$P(z) \;=\; \prod_{j=1}^{p}(z-a_j),$$

where the $a_j$ are distinct. Suppose that $f$ and $g$ are entire meromorphic functions that satisfy the identity

$$f(z)^2 \;=\; P(g(z)) \;=\; \prod_{j=1}^{p}(g(z)-a_j). \tag{9.4.8}$$

Every zero of $f^2$ has multiplicity at least two, so every zero of each $g(z)-a_j$ must be multiple. In other words, the $a_j$ are completely ramified values for $g$. It follows from Theorem 9.4.3 that the degree $p$ is at most 4, so the genus is at most 1. □

**Remarks.** Picard's theorem applies to every algebraic curve of genus $g\ge 2$, but the proof is more difficult. Such curves can be parametrized by automorphic functions; see Section 7.8.

## 9.5 Further properties of meromorphic functions

In this section we mention a number of results that are closely related to what has been covered in the chapter. Most of the details are left as exercises.

The second fundamental theorem is usually stated in a different form, with more stringent limitations on the exceptional set.

**Theorem 9.5.1.** *Suppose that $f$ is an entire meromorphic function and $k\ge 0$. For distinct points $a_1,a_2,\dots a_q$ in $\mathbb{S}$, the inequality*

$$\sum_{j=1}^{q} m(r,a_j,f) \;<\; 2T(r,f)-N_1(r,f)+O[\log r+\log T^\circ(r,f)]+O(\log r) \tag{9.5.1}$$

*holds for each $r$ outside an open set $\Delta$ such that $\int_\Delta r^k\,dr<\infty$.*

The proof is a minor adaptation of the proof of Theorem 9.3.3, then carried over to the Nevanlinna characteristic; see Exercise 18.

One application of this more stringent condition on the exceptional set $\Delta$ comes in connection with meromorphic functions of finite order. A function $f$ is said to have *order* $\rho$ if

$$\limsup_{r\to\infty} \frac{\log T(r,f)}{\log r} = \rho. \tag{9.5.2}$$

**Theorem 9.5.2.** *Suppose that the entire meromorphic function $f$ has finite order. Then the inequality (9.5.1) is valid for every $r$.*

For the proof, see Exercise 19.

## Exercises

Throughout these exercises, the functions considered are taken to be entire, meromorphic, and not constant.

1. Suppose that $f$ is meromorphic for $|z| < r+\delta$, $\delta > 0$, and has no zeros or poles with $|z| = r$. Prove that the change in $\arg f(re^{i\theta})$ from $\theta = 0$ to $\theta = 2\pi$ is the number of zeros of $f$ with $|z| < r$ minus the number of poles of $f$ with $|z| < r$. (Hint: write $f$ as
$$f(z) = \frac{\prod_{j=1}^{m}(z-a_j)}{\prod_{k=1}^{n}(z-b_k)}\, g(z),$$
where $g$ has no zeros or poles with $|z| \le r$.)
2. Suppose that $P$ is a polynomial. (a) Use the asymptotic behavior of $|P|$ as $|z| \to \infty$ to show that $P$ has only finitely many zeros.

   (b) Use (9.1.16) to show that $P$ has exactly $\deg P$ zeros, counting multiplicity.
3. Suppose that $P$ and $Q$ are polynomials with no common zeros. Use (9.1.16) to show that the rational function $P/Q$, as a function on the Riemann sphere $\mathbb{S}$, attains each value in $\mathbb{S}$ exactly $\max\{\deg P, \deg Q\}$ times.
4. (a) Deduce from (9.1.11) that the function $f(z) = e^z$ has no zeros.

   (b) For $a \neq 0, \infty$, determine $N(r,a,f)$ and $T(r,f)$.
5. Discuss $f(z) = \tan z$.
6. Prove Proposition 9.1.2.
7. Prove that for positive $x_1, x_2, \ldots, x_p$
$$\log^+\left(\sum_{j=1}^{p} x_j\right) \le \log p + \sum_{j=1}^{p} \log^+ x_j.$$
(Hint: write $x_j = 1+\varepsilon_j$.)
8. Prove that for positive $x_1, x_2, \ldots, x_p$,
$$\log^+(x_1 x_2 \cdots x_p) \le \sum_{j=1}^{p} \log^+ x_j.$$
9. Show that $T(r,0,fg) \le T(r,0,f) + T(r,0,g)$.

10. Show that if $k \neq 0$, then

$$|T(r,kf) - T(r,f)| \leq |\log|k||.$$

11. Show that

$$|T(r,f-a) - T(r,f)| \leq \log^+|a| + \log 2.$$

12. Show that if $f(0) \neq 0$ then

$$T(r,f) - T\left(r,\frac{1}{f}\right) = \log|f(0)|.$$

13. Prove that

$$\lim_{r\to\infty} \frac{T(r,f)}{\log r} < 2$$

if and only if $f$ is a linear fractional transformation.

14. Suppose $g$ is obtained from $f$ by a linear fractional transformation:

$$g(z) = \frac{af(z)+b}{cf(z)+d}, \quad ad-bc = 1.$$

Show that $|T(r,g) - T(r,f)|$ is a bounded function of $r$.

15. Show that

$$m_\rho(r,f) = \int_{\mathbb{C}} [m^\circ(r,a,f) - \log d(c(a),a)]\,\rho(a)\,dm(a),$$

where $c(a)$ is the leading coefficient in the Laurent expansion of $f(z)-a$ at $z=0$.

16. Suppose $g : (0,\infty) \to (0,\infty)$ and $g'' \leq 0$.

(a) Prove that $g$ is convex downward: the graph of $g(x)$, $0 < b < x < c$ lies above the line segment from $g(b)$ to $g(c)$.

(b) Prove that for any choice of positive $a_1,\ldots,a_n$ with $\sum_{j=1}^n a_j = 1$ and any positive $x_1,\ldots x_n$,

$$g(\sum_{j=1}^n a_j x_j) \geq \sum_{j=1}^n a_j g(x_j).$$

(In fact the case $n=2$ is part (a); prove the general case by induction.) This is a discrete form of Jensen's inequality.

(c) Prove a continuous form of Jensen's inequality. If $g$ is convex downward on an interval $[a,b]$, then

$$g\left(\frac{1}{b-a}\int_a^b f(x)\,dx\right) \geq \frac{1}{b-a}\int_a^b g\circ f(x)\,dx.$$

17. Prove that Theorem 9.4.1 is equivalent to: a non-constant entire function can omit at most one value.

18. Prove Theorem 9.5.1: replace the conditions (9.3.26) and (9.3.28) by: $\Delta_1$ is the set of $r$ such that
$$\lambda(r,f) \;\geq\; r^{k-1}L(r,f)^2$$
and $\Delta_2$ is the set of $r$ such that
$$L(r,f) \;\geq\; r^{k+1}K(r,f)^2.$$

19. Prove Theorem 9.5.2 by choosing $k$ in Theorem 9.5.1 to be larger than $\rho$, and noting that
$$(q-2)T(r,f)+N_1 \;<\; \sum_{j=1}^{q} N(r,a_j,f)+C_k\log r \tag{9.5.3}$$
for $r$ not in the exceptional set $\Delta$. Show that if $r\in\Delta$ is sufficiently large, then there is some $r^*$ not in $\Delta$ such that $r<r^*<r+r^{-k}$. Use this to show that (9.5.3) holds for each such $r$ with a larger (but fixed) choice of the constant $C_k$.
20. Suppose that $f$ is meromorphic in the disk $\{z:|z|<R\}$. Prove a version of Theorem 9.5.1 with $O(\log r)$ replaced by $-(k+\varepsilon)\log(R-r)]$, valid outside an exceptional set $\Delta$ with
$$\int_\Delta \frac{1}{(R-r)^k}\,dr \;<\; \infty.$$

## Remarks and further reading

Our presentations of the proof by Ahlfors of the second fundamental theorem, and of its applications, follow Hille [64], §14.6 and §14.7, (with some gaps filled and minor errors corrected).

There is a large literature on further developments of the Nevanlinna theory. These include refinements of the general theory, extensions to functions meromorphic in a disk (see Exercise 20) or in a sector, and to maps from one Riemann surface to another.

The early history of Nevanlinna's theory is nicely sketched by Gårding [47]. Hayman [61] is a very readable treatment of the first four decades of the theory. Goldberg and Ostrovskii [50] contains further results, and an appendix by Eremenko and Langley brings references into the 21st century. See also Rubel [124], Yang [146], and Zheng [148].

For applications to complex dynamics, see Bergweiler [19]. Some surprising connections with diophantine approximation were discovered by Vojta and others in the 1980s; see Cherry and Ye [32], Hu and Yang [67], and Ru [123].

# Chapter 10
# The gamma and beta functions



The factorial function is defined for positive integers $n$ by $n! = n(n-1)(n-2)\cdots 1$. It is convenient to shift by 1 and define the *gamma function* on the positive integers by

$$\Gamma(1) = 1, \quad \Gamma(n) = (n-1)!, \qquad n = 2,3,4,\dots .$$

Then $\Gamma$ satisfies the *functional equation*

$$\Gamma(n+1) = n\Gamma(n), \qquad n = 1,2,3,\dots , \tag{10.0.1}$$

and is uniquely determined by this, together with the condition $\Gamma(1) = 1$.

A number of mathematicians in the 17th and 18th centuries considered the problem of extending $\Gamma(x)$ to all positive values $x$, or, eventually, as a meromorphic function on $\mathbb{C}$, while preserving the functional equation:

$$\Gamma(z+1) = z\Gamma(z) \tag{10.0.2}$$

for each $z$ for which $\Gamma(z)$ is defined.

In this chapter we present two solutions to this problem, both due to Euler, and prove that they are equivalent. Examination of the product of gamma functions leads to the beta function, also due to Euler. Both functions occur naturally in many computations in analysis, probability, and statistical mechanics. Further results include Legendre's duplication formula, Euler's reflection formula, and asymptotic estimates of Stirling and Stieltjes.

## 10.1 Euler's product solution

Euler found a solution to the problem of extending the factorial function, $\Gamma(n) = (n-1)!$, based on the identity

$$\Gamma(n+k) = 1\cdot 2\cdots(n-1)\cdot[n(n+1)\cdots(n+k-1)] = \Gamma(n)\,(n)_k, \tag{10.1.1}$$



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_10

where $(n)_k$ denotes the *shifted factorial* or *Pochhammer symbol*, defined for any $z \in \mathbb{C}$ by

$$(z)_k = z(z+1)\cdots(z+k-1), \qquad z \in \mathbb{C}.$$

Now $\Gamma(n+k)$ is symmetric in $k$ and $n$, leading to the identity

$$\Gamma(k) = \frac{\Gamma(n)\,(n)_k}{(k)_n}.$$

Euler's idea was to let $n \to \infty$ with $k$ fixed, then $(n)_k \sim n^k$, so

$$\Gamma(k) = \lim_{n\to\infty} \frac{\Gamma(n)\,n^k}{(k)_n}.$$

Replacing $k$ by $z$ leads to a possible general definition:

$$\Gamma(z) = \lim_{n\to\infty} \frac{\Gamma(n)\,n^z}{(z)_n}. \tag{10.1.2}$$

To check convergence we use the two identities

$$\begin{aligned}\frac{(z)_n}{\Gamma(n)} &= z \cdot \frac{z+1}{1} \cdot \frac{z+2}{2} \cdots \frac{z+n-1}{n-1} \\ &= z \cdot \left(1+\frac{z}{1}\right)\left(1+\frac{z}{2}\right)\cdots\left(1+\frac{z}{n-1}\right)\end{aligned}$$

and

$$n = \frac{\Gamma(n+1)}{\Gamma(n)} = \frac{2}{1}\cdot\frac{3}{2}\cdots\frac{n}{n-1} = (1+1)\left(1+\frac{1}{2}\right)\cdots\left(1+\frac{1}{n-1}\right).$$

Therefore

$$\frac{\Gamma(n)\,n^z}{(z)_n} = \frac{1}{z}\prod_{j=1}^{n-1}\left(1+\frac{z}{j}\right)^{-1}\left(1+\frac{1}{j}\right)^{z}. \tag{10.1.3}$$

The logarithm of the $j$-th factor is $O(j^{-2})$ for large $j$, see Lemma 1.5.1. Therefore we have convergence:

$$\Gamma(z) = \frac{1}{z}\prod_{j=1}^{\infty}\left(1+\frac{z}{j}\right)^{-1}\left(1+\frac{1}{j}\right)^{z}, \tag{10.1.4}$$

so long as no factor blows up, i.e. so long as $z$ is not a non-positive integer (see the discussion in Section 1.5). It is easily checked from this product that the poles are simple and that there are no zeros. We have

**Theorem 10.1.1.** *$\Gamma$ is meromorphic in $\mathbb{C}$ with simple poles at the non-positive integers, and no zeros.*

It can be checked that $\Gamma$ satisfies the functional equation (10.0.2): Exercise 1.

Let us look at the reciprocal:

$$
\begin{aligned}
\frac{(z)_n}{\Gamma(n)\,n^z} &= zn^{-z}\prod_{j=1}^{n-1}\frac{z+j}{j} \;=\; zn^{-z}\prod_{k=1}^{n-1}\left(1+\frac{z}{j}\right)\\
&= zn^{-z}\prod_{j=1}^{n-1}\left[\left(1+\frac{z}{j}\right)e^{-z/j}\right]\exp\left(z+\frac{z}{2}+\cdots+\frac{z}{n-1}\right)\\
&= z\exp\left\{z\left(1+\frac{1}{2}+\cdots+\frac{1}{n-1}-\log n\right)\right\}\prod_{j=1}^{n-1}\left(1+\frac{z}{j}\right)e^{-z/j}.
\end{aligned}
$$

The logarithms of the factors in the product are $O(j^{-2})$, while the coefficient of $z$ in the exponential factor on the left is

$$
\sum_{j=1}^{n-1}\frac{1}{j}-\log n \;=\; \sum_{j=1}^{n-1}\frac{1}{j}-\int_1^n\frac{dt}{t} \;=\; \sum_{j=1}^{n-1}\int_j^{j+1}\left\{\frac{1}{j}-\frac{1}{t}\right\}dt.
$$

The integrand of $j$-th integral is $O(j^{-2})$, so the series of integrals converges. The limit is, by definition, *Euler's constant*

$$
\gamma = \lim_{n\to\infty}\left[\sum_{j=1}^{n-1}\frac{1}{j}-\log n\right] \;=\; 0.5772\ldots\,. \tag{10.1.5}
$$

For later use, we note that the preceding argument shows that

$$
-\gamma+\sum_{k=1}^{n-1}\frac{1}{k} \;=\; \log n+O(n^{-1}). \tag{10.1.6}
$$

**Theorem 10.1.2.** *The reciprocal of the gamma function is the product*

$$
\frac{1}{\Gamma(z)} \;=\; ze^{\gamma z}\prod_{n=1}^{\infty}\left(1+\frac{z}{n}\right)e^{-z/n}. \tag{10.1.7}
$$

*It is an entire function with simple zeros at the non-positive integers.*

*Proof:* The previous argument established the product representation (10.1.7). The partial products are entire functions of $z$ and the convergence is uniform on bounded sets, so the product is entire (Proposition 1.2.6). The zeros are precisely the zeros of the factors. □

Consider now the product of reciprocals

$$
s(z) \;=\; \frac{1}{\Gamma(z)\,\Gamma(1-z)}. \tag{10.1.8}
$$

The function $s$ is an entire function of $z$. The zeros of $s$ are precisely the integers. Moreover $s(-z) = -s(z)$ and $s(z+1) = -s(z)$. (The proof of these statements is left

as an exercise.) This suggests that $s(z)$ may be a fixed multiple of $\sin \pi z$. To prove this we use a second, more commonly used, representation of the gamma function.

## 10.2 Euler's integral solution and the beta function

Euler noted that

$$\int_0^1 (-\log x)^n \, dx = n!, \qquad n = 0, 1, 2, \dots . \tag{10.2.1}$$

This suggests defining $\Gamma$ by

$$\Gamma(z) = \int_0^1 (-\log x)^{z-1} \, dx.$$

The integral converges for $\operatorname{Re} z > 0$. A change of variables $x = e^{-t}$ leads to the more commonly used form

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} \, dt, \qquad \operatorname{Re} z > 0. \tag{10.2.2}$$

We will see below that this is consistent with (10.1.4). One can derive directly that $\Gamma$, as defined by (10.2.2), has an extension that is meromorphic in all of $\mathbb{C}$, with simple poles at the non-negative integers: Exercises 8 and 9.

In order to prove equivalence of (10.1.4) and (10.2.2), we make an excursion into products and the beta function. With the definition (10.2.2), and assuming that $\operatorname{Re} a > 0$, $\operatorname{Re} b > 0$,

$$\Gamma(a)\Gamma(b) = \int_0^\infty \int_0^\infty e^{-(s+t)} s^{a-1} t^{b-1} \, ds \, dt. \tag{10.2.3}$$

After some changes of variables, (10.2.3) becomes

$$\Gamma(a)\Gamma(b) = \Gamma(a+b) \int_0^\infty u^{a-1} (1+u)^{-a-b} \, du; \tag{10.2.4}$$

see Exercise 10. (For positive integer values of $a$ and $b$, an equivalent formula was known to John Wallis about 1655.)

The integral here gives, by definition, one form of the *beta function*:

$$\mathrm{B}(a,b) = \int_0^\infty u^{a-1} (1+u)^{-a-b} \, du, \qquad \operatorname{Re} a > 0, \ \operatorname{Re} b > 0. \tag{10.2.5}$$

The change of variables $u = 1/(1-s)$ leads to a second, more commonly used form

$$\mathrm{B}(a,b) = \int_0^1 s^{a-1} (1-s)^{b-1} \, ds, \qquad \operatorname{Re} a > 0, \ \operatorname{Re} b > 0. \tag{10.2.6}$$

Rewriting (10.2.4) as

$$\mathrm{B}(a,b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)} \tag{10.2.7}$$

shows that B can be extended so as to be defined and holomorphic in $a$ and in $b$, for any pair $a,b$ for which neither $a$ nor $b$ is a pole of $\Gamma$.

We will use (10.2.6) and (10.2.7) to prove that the two definitions of $\Gamma$ agree.

**Lemma 10.2.1.** *As defined by (10.2.2), $\Gamma$ satisfies*

$$\frac{\Gamma(x)\,x^z}{\Gamma(x+z)} \to 1 \quad \text{as} \quad x \to +\infty, \qquad \mathrm{Re}\, z > 0.$$

*Proof:* With $\Gamma$ defined by (10.2.2), the left side here is $x^z B(z,x)/\Gamma(z)$, so we want to prove

$$x^z \int_0^1 s^{z-1}\,(1-s)^{x-1}\,ds \to \Gamma(z) \quad \text{as} \quad x \to +\infty. \tag{10.2.8}$$

We write this expression as

$$x^z \int_0^{1/2} s^{z-1}\,(1-s)^{x-1}\,ds + x^z \int_{1/2}^1 s^{z-1}\,(1-s)^{x-1}\,ds.$$

The integrand in the second integral is dominated by $2^{2-x}$, so the second integral decays exponentially in $x$ and its product with $x^z$ has limit zero. The first summand becomes, after letting $s = t/x$,

$$\int_0^{x/2} t^{z-1} \left(1 - \frac{t}{x}\right)^x \left(1 - \frac{t}{x}\right)^{-1} dt.$$

The integrand is dominated by

$$t^{\mathrm{Re}\, z-1} \cdot e^{-t} \cdot 2,$$

and, as $x \to \infty$, it converges uniformly on any given bounded interval to $t^{z-1}e^{-t}$. The limit (10.2.8) follows. □

**Theorem 10.2.2.** *The two definitions (10.1.2) and (10.2.2) agree, for* $\mathrm{Re}\, z > 0$.

*Proof:* Note that the functional equation (10.0.2) can be extended to show that

$$\Gamma(z+k) = (z)_k\,\Gamma(z). \tag{10.2.9}$$

Thus, with $\Gamma$ defined by the integral formula (10.2.2), we can use (10.2.8) to obtain

$$\frac{\Gamma(n)\,n^z}{(z)_n} = \frac{\Gamma(n)\,n^z}{\Gamma(n+z)} \cdot \Gamma(z) \to \Gamma(z)$$

as $n \to \infty$. Thus the two definitions coincide for $\mathrm{Re}\, z > 0$. □

It follows that the analytic continuation of the integral form (10.2.2) outlined in Exercises 8 and 9 agrees with the product form (10.1.4) everywhere.

## 10.3 Legendre's duplication formula

For positive integers $n$,

$$\begin{aligned}\Gamma(2n) &= (2n-1)\,! \\ &= 1\cdot 3\cdots(2n-1)\cdot 2\cdot 4\cdots(2n-2) \\ &= 2^{n-1}\left[\left(\frac{1}{2}\right)\cdot\left(\frac{3}{2}\right)\cdots\left(\frac{1}{2}+n-1\right)\right]\cdot 2^n\cdot(n-1)\,!.\end{aligned}$$

Invoking the functional equation (10.2.9), we get

$$\Gamma(\tfrac{1}{2})\cdot\Gamma(2n) \;=\; 2^{2n-1}\Gamma(n+\tfrac{1}{2})\Gamma(n). \tag{10.3.1}$$

Since $\Gamma(1/2)=\pi^{1/2}$ (Exercise 20), (10.3.1) can also be written as

$$\Gamma(2n) \;=\; \pi^{-1/2}\,2^{2n-1}\Gamma(n+\tfrac{1}{2})\Gamma(n).$$

This is the integer case of Legendre's *duplication formula* for the gamma function.

**Theorem 10.3.1.** (Legendre) *For any* $2z\neq 0,-1,-2,-3,\ldots,$

$$\Gamma(\tfrac{1}{2})\Gamma(2z) \;=\; 2^{2z-1}\Gamma(z+\tfrac{1}{2})\Gamma(z). \tag{10.3.2}$$

For a proof, see Exercise 15 or Exercise 16.

## 10.4 The reflection formula and the product formula for sine

Let us return to the function $s(z)=1/\Gamma(1-z)\Gamma(z)$.

**Theorem 10.4.1.** (Euler's reflection formula)

$$\Gamma(z)\Gamma(1-z) \;=\; \frac{\pi}{\sin\pi z}, \qquad z\notin\mathbb{Z}. \tag{10.4.1}$$

*Proof:* It is enough to prove this under the assumption that $0<\operatorname{Re}z<1$. We start with the identity

$$\Gamma(z)\Gamma(1-z) \;=\; \mathrm{B}(z,1-z) \;=\; \int_0^\infty \frac{t^{z-1}}{1+t}\,dt. \tag{10.4.2}$$

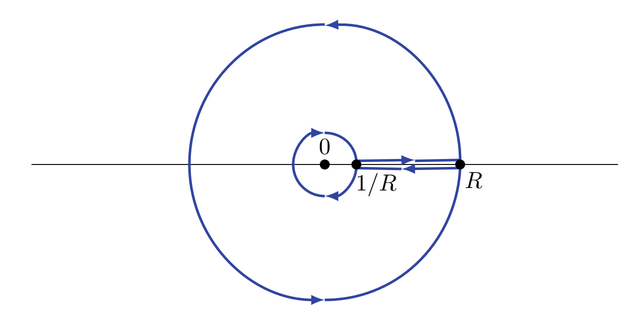



**Fig. 10.1** The contour $\Gamma_R$

Given $R > 1$, let $\Gamma_R$ denote the contour that runs along the real axis from $R$ to $1/R$, in the clockwise direction around the circle of radius $1/R$ centered at the origin, then along the real axis from $1/R$ to $R$, and counterclockwise around the circle of radius $R$ centered at the origin, back to $R$; see Figure 10.1.

For $t$ in the domain enclosed by this contour, we take the branch of $\log t$ that has a real limit as $t$ approaches the positive real axis from the upper half plane. Then the limit as $t$ approaches the positive real axis from the lower half plane differs by a factor $2\pi i$. The corresponding determinations of $t^z = \exp(z \log t)$ differ by a multiplicative factor $\exp(2\pi i z)$.

The function $f(t) = t^{z-1}/(1+t)$ has a unique pole in the domain enclosed by $\Gamma_R$, at $t = -1$, with residue $-\exp(i\pi z)$. Therefore

$$\frac{1}{2\pi i} \int_{\Gamma_R} \frac{t^{z-1}}{1+t}\, dt \;=\; -e^{i\pi z}.$$

For $|t| = R$ the integrand is $O(R^{\mathrm{Re}\, z-2})$, so the integral over this circle has limit zero as $R \to \infty$. It follows from these considerations that $\mathrm{B}(z, 1-z)$ satisfies

$$\frac{1 - e^{2\pi i}}{2\pi i}\, \mathrm{B}(z, 1-z) \;=\; -e^{i\pi z}.$$

Together with (10.4.2), this is (10.4.1). □

The reflection formula (10.4.1) can be used, together with (10.1.4) and (10.0.2), to prove *Euler's product formula for sine*:

$$\sin \pi z \;=\; \pi z \prod_{n=1}^{\infty} \left(1 - \frac{z^2}{n^2}\right); \tag{10.4.3}$$

see Exercise 17. This in turn can be used to derive the result that first made Euler famous (among European mathematicians):

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \;=\; \frac{\pi^2}{6}; \tag{10.4.4}$$

see Exercise 18.

## 10.5 Asymptotics of the gamma function

The most widely used result about the behavior of the gamma function for large values of the argument is *Stirling's approximation*:

**Theorem 10.5.1.** *The $\Gamma$ function satisfies*

$$\Gamma(x) = \left(\frac{x}{e}\right)^x \left[\left(\frac{2\pi}{x}\right)^{1/2} + O(x^{-3/2})\right] \tag{10.5.1}$$

*as $x \to +\infty$.*

*Proof:* Starting with

$$\Gamma(x) = \frac{1}{x}\Gamma(x+1) = \frac{1}{x}\int_0^\infty e^{-t}t^x\,dt,$$

note that the integrand has a maximum at $t = x$, which suggests a change of variables $t = xu$:

$$\Gamma(x) = x^x \int_0^\infty (ue^{-u})^x\,du = \left(\frac{x}{e}\right)^x \int_0^\infty \left(ue^{1-u}\right)^x du.$$

The integrand decays exponentially with respect to $x$ away from $u = 1$, so we may restrict attention to a small interval centered at $u = 1$. The function $ue^{1-u}$ agrees to second order at $u = 1$ with the function $e^{-(u-1)^2/2}$, so in a neighborhood of $u = 1$,

$$ue^{1-u} = e^{-(u-1)^2/2}\left[1 + O\left((u-1)^3\right)\right].$$

Because of the exponential decay, the asymptotics of $\Gamma(x)$ as $x \to \infty$ agree with those of

$$\left(\frac{x}{e}\right)^x \int_{-\infty}^\infty e^{-xs^2/2}[1 + O(xs^3)]\,ds. \tag{10.5.2}$$

Note that $\int_{-\infty}^\infty \exp(-xs^2/2)s^3\,ds = 0$, so (10.5.2) is

$$\left(\frac{x}{e}\right)^x \int_{-\infty}^\infty e^{-xs^2/2}[1 + O(xs^3)]\,ds = \left(\frac{x}{e}\right)^x \left[\left(\frac{2\pi}{x}\right)^{1/2} + O(x^{-3/2})\right]. \qquad \square$$

Here we used the well-known evaluation

$$\begin{aligned}\left[\int_{-\infty}^\infty e^{-x^2/2}\,dx\right]^2 &= \iint_{\mathbb{R}^2} e^{-(x^2+y^2)/2}\,dx\,dy \\ &= \int_0^\infty \int_0^{2\pi} e^{-r^2/2} r\,d\theta\,dr \\ &= 2\pi \int_0^\infty e^{-r^2/2}\,d(r^2/2) = 2\pi.\end{aligned}$$

The approximation (10.5.1) was originally proved for integer $x$; see the discussion in Chapter 24 of [121]. It plays an important role in statistical mechanics, for example. In Chapter 13 we will need a more global asymptotic estimate. To make the strategy of the proof more transparent we start with another derivation of the integer case. The idea is to estimate $\log\Gamma(n)$, and the first observation is that it is easier to deal with integrals than sums:

$$\begin{aligned}\log\Gamma(n) &= \sum_{k=1}^{n-1}\log k \;=\; \sum_{k=1}^{n-1}\int_1^k \frac{ds}{s}\\ &= (n-1)\int_1^2 \frac{ds}{s} + (n-2)\int_2^3 \frac{ds}{s} + \cdots + \int_{n-2}^{n-1}\frac{ds}{s}\\ &= \int_1^{n-1}\frac{n-1-[s]}{s}\,ds, \quad [s] \;=\; \text{largest integer } \le s.\end{aligned}$$

The next step is to approximate the step function $[s]$ by the continuous function $s-\frac{1}{2}$, so that

$$\begin{aligned}\log\Gamma(n) &= \int_1^{n-1}\frac{n-s-\frac{1}{2}}{s}\,ds + \int_1^{n-1}\frac{s-[s]-\frac{1}{2}}{s}\,ds\\ &= \left(n-\tfrac{1}{2}\right)\log(n-1)-(n-2)+\int_1^{n-1}\frac{s-[s]-\frac{1}{2}}{s}\,ds. \qquad (10.5.3)\end{aligned}$$

Writing $\log(n-1) \;=\; \log n(1-1/n)$, we see that

$$\left(n-\tfrac{1}{2}\right)\log(n-1) \;=\; \left(n-\tfrac{1}{2}\right)\log n - 1 + O(n^{-1}). \qquad (10.5.4)$$

The second integral in (10.5.3) is to be considered as a remainder. The point of using $s-\frac{1}{2}$ rather than $s$ as an approximation to $[s]$ is that $s-[s]\ge 0$. Therefore there is no cancellation, and the resulting integral against $1/s$ is $O(\log n)$. However $s-[s]-\frac{1}{2}$ has integral 0 over each interval $[k,k+1]$, so the extra term is smaller. In fact the function

$$f(s) \;=\; \int_0^s\left(s-[s]-\frac{1}{2}\right)ds \qquad (10.5.5)$$

is periodic, since the integrand is periodic and $f(k+1)=f(k)=0$. Therefore $f$ is bounded. Integration by parts gives

$$\int_1^{n-1}\frac{s-[s]-\frac{1}{2}}{s}\,ds \;=\; \int_1^{n-1}\frac{f(s)}{s^2}\,ds \;=\; \int_1^{\infty}\frac{f(s)}{s^2}\,ds + O(n^{-1}). \qquad (10.5.6)$$

From (10.5.3), (10.5.4), (10.5.5), and (10.5.6) we obtain

$$\log\Gamma(n) \;=\; \left(n-\tfrac{1}{2}\right)\log n - n + C + O(n^{-1}), \qquad (10.5.7)$$

where $C$ is constant. According to (10.5.1), $C=\log\sqrt{2\pi}$. (For an independent determination of $C$, see Exercise 21.)

Since $\Gamma(z)$ has poles at the non-positive integers, the approximation (10.5.1) fails near the negative real axis. Starting from (10.5.7), and adapting the argument just given, we shall see that the approximation extends to the complement of any angular sector that includes the negative real axis.

**Theorem 10.5.2.** (Stieltjes) *The approximation*

$$\Gamma(z) \;=\; \left(\frac{z}{e}\right)^z \left[\left(\frac{2\pi}{z}\right)^{1/2} + O(z^{-3/2})\right] \tag{10.5.8}$$

*is valid as $z \to \infty$, uniformly in any sector $\{z : |\arg z| \le \pi - \delta\}$, $\delta > 0$.*

*Proof:* We work with the logarithm of the product representation of $1/\Gamma(z)$, (10.1.7):

$$\begin{aligned}
\log\Gamma(z) &= -\log z - \gamma z + \sum_{n=1}^{\infty}\left\{\frac{z}{n} - \log\left(1+\frac{z}{n}\right)\right\} \\
&= -\log z + z\left(\sum_{k=1}^{n-1}\frac{1}{k} - \gamma\right) - \sum_{k=1}^{n-1}\log\left(1+\frac{z}{k}\right) + O(n^{-1}).
\end{aligned}$$

Taking into account (10.1.5) and writing $1+z/k$ as $(z+k)/k$, we have

$$\begin{aligned}
\log\Gamma(z) &= -\log z + z\log n - \sum_{k=1}^{n-1}\log(k+z) + \sum_{k=1}^{n-1}\log k + O(n^{-1}) \\
&= -\log z + z\log n - \sum_{k=1}^{n-1}\log(k+z) + \log\Gamma(n) + O(n^{-1}). \quad (10.5.9)
\end{aligned}$$

Adapting the previous argument,

$$\begin{aligned}
\sum_{k=1}^{n-1}\log(z+k) &= \sum_{k=1}^{n-1}\int_1^{k+z}\frac{ds}{s} \\
&= \sum_{k=1}^{n-1}\left\{\int_z^{k+z}\frac{ds}{s} + \int_1^z\frac{ds}{s}\right\} \;=\; \sum_{k=1}^{n-1}\left\{\int_0^k\frac{ds}{s+z}\right\} + (n-1)\log z \\
&= \int_0^{n-1}\frac{n-1-[s]}{s+z}\,ds + (n-1)\log z. \quad (10.5.10)
\end{aligned}$$

Next,

$$\int_0^{n-1}\frac{n-1-[s]}{s+z}\,ds \;=\; \int_0^{n-1}\frac{n-\frac{1}{2}-s}{z+s}\,ds + \int_0^{n-1}\frac{s-\frac{1}{2}-[s]}{z+s}\,ds. \tag{10.5.11}$$

The first integral on the right is

$$\begin{aligned}
&\int_0^{n-1} \frac{(n-\frac{1}{2}+z)-(z+s)}{z+s}\,ds \\
&\quad = (n-\tfrac{1}{2}+z)[\log(n-1+z)-\log z]-(n-1) \\
&\quad = (n-\tfrac{1}{2}+z)\left[\log n+\log\left(1+\frac{z-1}{n}\right)-\log z\right]-(n-1) \\
&\quad = (n-\tfrac{1}{2}+z)[\log n-\log z]+(z-1)-(n-1)+O(n^{-1}). \qquad (10.5.12)
\end{aligned}$$

Consider now the second integral on the right in (10.5.11):

$$\int_0^{n-1} \frac{s-\frac{1}{2}-[s]}{z+s}\,ds \;=\; \int_0^{n-1} \frac{f(s)}{(z+s)^2}\,ds, \tag{10.5.13}$$

where $f$ is given as before, by (10.5.5). For $s \geq 0$ and $|\theta| = |\arg z| < \pi-\delta$, $\delta > 0$, it follows that $\cos\theta \geq -1+\varepsilon$ for some $\varepsilon > 0$, so

$$|z+s|^2 \;=\; |z|^2+2|z|s\cos\theta+s^2 \;\geq\; (|z|^2+s^2)\min\{1,1+\cos\theta\} \;\geq\; \varepsilon(|z|^2+s^2).$$

Therefore the integral (10.5.13) is $O(|z|^{-1})$, uniformly in the sector. Combining this with (10.5.7) to (10.5.13), we have

$$\begin{aligned}
\log\Gamma(z) &= -\log z+z\log n+\log\Gamma(n) \\
&\qquad -\left\{(n-\tfrac{1}{2}+z)[\log n-\log z]+z-n+(n-1)\log z\right\}+O(n^{-1}+|z|^{-1}) \\
&= (z-\tfrac{1}{2})\log z-z+[\log\Gamma(n)-(n-\tfrac{1}{2})\log n+n]+O(n^{-1}+|z|^{-1}) \\
&= (z-\tfrac{1}{2})\log z-z+\log\sqrt{2\pi}+O(n^{-1}+|z|^{-1}).
\end{aligned}$$

Letting $n \to \infty$, we obtain (10.5.8). □

## Exercises

1. Using (10.1.4), verify the functional equation (10.0.2).
2. Verify the following properties of the product (10.1.8): it is entire, odd, and satisfies $s(z+1) = -s(z)$.
3. Use (10.1.4) to prove *Wallis's formula*

$$\frac{\pi}{4} \;=\; \frac{2}{3}\cdot\frac{4}{3}\cdot\frac{4}{5}\cdot\frac{6}{5}\cdots\frac{2n}{2n+1}\cdot\frac{2n+2}{2n+1}\cdots.$$

4. Show that if $x$ and $y$ are real and $x \neq 0,-1,-2,\ldots$, then

$$\left|\frac{\Gamma(x)}{\Gamma(x+iy)}\right|^2 \;=\; \prod_{n=0}^{\infty}\left\{1+\frac{y^2}{(x+n)^2}\right\},$$

proving that $|\Gamma(x+iy)| < |\Gamma(x)|$ unless $y = 0$.

5. Prove (10.2.1).
6. Suppose $\Gamma$ has a representation of the form

$$\Gamma(z) \;=\; \int_0^\infty e^{-t} f(t,z)\,dt.$$

Show that the functional equation will hold, for $\operatorname{Re} z > 0$, provided that

$$\frac{\partial}{\partial t}[f(t,z+1)] \;=\; zf(t,z).$$

The simplest choice here is $f(z,t) \;=\; t^{z-1}$, leading to the formula (10.2.2).
7. Verify that (10.2.2) implies that for positive integer $n$, $\Gamma(n) = (n-1)\,!$.
8. Break the interval of integration in (10.2.2) into $I_0 = (0,1)$ and $I_1 = (1,\infty)$. Show that the $I_1$ piece extends to an entire function of $z$.
9. In the $I_0$ piece from the previous exercise, expand the exponential in power series and integrate term-by-term to show that this part extends as an entire meromorphic function with simple poles, and show that the residue at $-k$ is $(-1)^k/k\,!$.
10. In (10.2.3), change variables to $u,t$ with $u = s/t$ and then to $u,x$ with $x = t(1+u)$ to derive (10.2.4).
11. Derive (10.2.6) from (10.2.5).
12. Show that if $\operatorname{Re}\alpha > 0$, $\operatorname{Re}\beta > 0$, and $u < x$ or $u > y$, then

$$\int_x^y \frac{(x-t)^{\alpha-1}(t-y)^{\beta-1}}{|t-u|^{\alpha+\beta}}\,dt \;=\; \mathrm{B}(\alpha,\beta)\,\frac{(x-y)^{\alpha+\beta-1}}{|x-u|^{\beta}\,|y-u|^{\alpha}}.$$

13. Show that if $\operatorname{Re} a > 0$ and $\operatorname{Re} b > 0$, then

$$\int_0^1 t^{a-1}(1-t^2)^{b-1}\,dt \;=\; \frac{1}{2}\,\mathrm{B}\left(\frac{1}{2}a, b\right).$$

14. Show that if $\operatorname{Re} a > 0$ and $\operatorname{Re} b > 0$, then

$$\int_0^{\pi/2} \sin^{a-1}\theta\,\cos^{b-1}\theta\,d\theta \;=\; \frac{1}{2}\,\mathrm{B}\left(\frac{1}{2}a, \frac{1}{2}b\right).$$

15. Prove the duplication formula (10.3.2). (Hint: by uniqueness of analytic continuation, it may be assumed that $\operatorname{Re} z > 0$. Start with

$$B(z,z) \;=\; 2\int_0^{1/2} s^{z-1}(1-s)^{z-1}\,ds$$

and let $t \;=\; 4s(1-s)$.)
16. Use (10.1.2), (10.2.1), and (10.3.1) to prove (10.3.2).
17. Use (10.4.1), (10.0.2), and (10.1.4) to prove (10.4.3).
18. Use (10.4.3) to prove (10.4.4).

19. Show that the second derivative of $\log \Gamma(z)$ is $\sum_n^\infty (z+n)^{-2}$. Thus $\log \Gamma(x)$ is a convex function for $x > 0$. The *Bohr–Mollerup Theorem* says that $\Gamma$ is the unique meromorphic extension of $(n-1)!$ that has this property.
20. Show that

$$\Gamma(\tfrac{1}{2}) \;=\; \sqrt{\pi}.$$

This can be done by a change of variables in either (10.2.2) or (10.2.7); in the case of (10.2.2), it reduces to evaluating $\int_0^\infty e^{-x^2}\,dx$.
21. Evaluate the constant in the approximation (10.5.8) by considering $\Gamma(z)$ for $z = \frac{1}{2} + it$, $t$ real, $t \to \infty$. Note that

$$|\Gamma(\tfrac{1}{2}+it)|^2 \;=\; \Gamma(\tfrac{1}{2}+it)\Gamma(\tfrac{1}{2}-it) \;=\; \frac{\pi}{\sin(\frac{1}{2}+it)\pi} \;=\; \frac{\pi}{\cosh(\pi t)}.$$

22. (a) Show that the approximation (10.5.8) can be sharpened to

$$\Gamma(z) \;=\; \left(\frac{z}{e}\right)^z \left(\frac{2\pi}{z}\right)^{1/2} \left[1 + \frac{1}{12z} + O(z^{-2})\right],$$

uniformly for $|\arg z| \le \pi - \delta$.

(b) Describe how to get further improvements to the approximation.

## Remarks and further reading

Artin's treatment of the gamma function [12] is classic. For the history, see Nielsen [108], Dutka [41], and Roy [121]. Two natural characterizations of the gamma function are due to Bohr and Mollerup [25] and to Wielandt: see [119] or [16], §2.4.

The gamma function and its properties are the basis for integral representations of solutions of hypergeometric equations and generalized hypergeometric equations; see Section 6.6 for the hypergeometric equation and the exposition [15] for the general case.

The gamma and beta functions provide many standard probability distributions; see any text on probability, statistics, or statistical mechanics. Additional identities and other information on the functions themselves can be found in any book on "special functions," e.g. [7], [16], [111].

# Chapter 11
# The Riemann zeta function




The *Riemann zeta function* is defined for $\mathrm{Re}\, s > 1$ by the series

$$\zeta(s) \;=\; 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \dots \,. \tag{11.0.1}$$

Euler investigated this series, long before Riemann, and noted that

$$\zeta(s) \;=\; \prod_{p \text{ prime}} \left( \frac{1}{1 - p^{-s}} \right). \tag{11.0.2}$$

This follows by expanding

$$\frac{1}{1 - p^{-s}} \;=\; 1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \frac{1}{p^{3s}} + \dots$$

and using the uniqueness of prime factorization: a positive integer $n$ has a unique representation

$$n \;=\; p_1^{d_1} p_2^{d_2} \dots p_m^{d_m},$$

where $p_1 < p_2 < \dots < p_m$ are primes.

As Euler noted, the fact that the series (11.0.1) diverges at $s = 1$ gives another proof that the set of primes is infinite—in fact $\sum_p (1/p)$ diverges. (This is only the simplest of the connections between properties of the zeta function and properties of primes.)

Convergence of the product (11.0.2) for $\mathrm{Re}\, s > 1$ and divergence of the sum $\sum_p (1/p)$ can be established by taking the logarithm of (11.0.2). In fact there is a quantitative estimate on the rate of divergence:

$$\sum_{p \le n} \frac{1}{p} \;\sim\; \log \log n \quad \text{as } n \to \infty. \tag{11.0.3}$$

See Exercises 6 and 7.



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_11

Further properties of the zeta function are developed in this chapter: its extension as an entire meromorphic function, evaluation at the even integers, a functional equation, and the fact that $\zeta(s) \neq 0$ when $\mathrm{Re}\, s = 1$. This last fact may seem to be a technical point, but it was a key step in the original proofs of the prime number theorem.

## 11.1 Properties of $\zeta$

It can be seen that $\zeta$ is holomorphic on the half plane $\{s : \mathrm{Re}\, s > 1\}$. This follows more easily from an integral representation, which itself follows from the identity

$$n^{-s} = \frac{1}{\Gamma(s)} \int_0^\infty e^{-nt}\, t^{s-1}\, dt, \qquad \mathrm{Re}\, s > 1.$$

(Take $nt$ as the variable of integration.) Summing,

$$\zeta(s) = \frac{1}{\Gamma(s)} \int_0^\infty \frac{e^{-t}}{1-e^{-t}}\, t^{s-1}\, dt = \frac{1}{\Gamma(s)} \int_0^\infty \frac{t^{s-1}\, dt}{e^t - 1}, \qquad \mathrm{Re}\, s > 1. \tag{11.1.1}$$

Let us write this in two pieces:

$$\zeta(s) = \zeta_0(s) + \zeta_1(s) = \frac{1}{\Gamma(s)} \int_0^1 \frac{t^{s-1}\, dt}{e^t - 1} + \frac{1}{\Gamma(s)} \int_1^\infty \frac{t^{s-1}\, dt}{e^t - 1}.$$

The function $\zeta_1$ extends as an entire function of $s$, with zeros at the poles of $\Gamma$, i.e. the non-positive integers. To study $\zeta_0$ we look more closely at the integrand. Now

$$\begin{aligned}
\frac{1}{e^t - 1} &= \frac{1}{t + \frac{1}{2}t^2 + O(t^3)} = \frac{1}{t} \cdot \frac{1}{1 + t/2 + O(t^2)} \\
&= \frac{1}{t} \cdot \big(1 - t/2 + O(t^2)\big) = \frac{1}{t} - \frac{1}{2} + O(t)
\end{aligned}$$

for $|t| < 2\pi$. We define a function $f$, holomorphic for $|t| < 2\pi$, by

$$f(t) = \frac{1}{e^t - 1} - \frac{1}{t} + \frac{1}{2}$$

and rewrite

$$\begin{aligned}
\zeta(s) &= \frac{1}{\Gamma(s)} \left\{ \int_0^1 \left[ \frac{1}{t} - \frac{1}{2} + f(t) \right] t^{s-1}\, dt + \int_1^\infty \frac{t^{s-1}}{e^t - 1}\, dt \right\} \\
&= \frac{1}{\Gamma(s)} \left\{ \frac{1}{s-1} - \frac{1}{2s} + \int_0^1 f(t)\, t^{s-1}\, dt + \int_1^\infty \frac{t^{s-1}}{e^t - 1}\, dt \right\}, 
\end{aligned} \tag{11.1.2}$$

for $\mathrm{Re}\, s > 1$. Since $f(t) = O(t)$, this formula can be used to extend $\zeta$ to the half-space $\{s : \mathrm{Re}\, s > 0\}$.

This process can be continued to the entire plane. It is an exercise to check that $f$ is holomorphic in the disk $\{z : |z| < 2\pi\}$ and is odd. Therefore $f$ has a power series expansion that converges uniformly on the interval $[0,1]$, with only odd powers:

$$\frac{1}{e^t - 1} - \frac{1}{t} + \frac{1}{2} \;=\; \sum_{m=1}^{\infty} \frac{B_{2m}}{(2m)!}\, t^{2m-1}. \tag{11.1.3}$$

The $B_{2m}$ are, by definition, the *Bernoulli numbers*. They can be computed recursively by multiplying (11.1.3) by $e^t - 1$. (There are a number of different conventions for numbering these coefficients and choosing their signs: let the reader beware.)

The integral defining $\zeta_0$ can be integrated term-by-term, giving

$$\zeta_0(s) \;=\; \frac{1}{\Gamma(s)} \left[ \frac{1}{s-1} - \frac{1}{2s} + \sum_{k=1}^{\infty} \left( \frac{B_{2k}}{(2k)!} \cdot \frac{1}{s+2k-1} \right) \right].$$

The expression in brackets has simple poles at $s = 1$ and $s = 0$, and at the odd negative integers $s = -(2k-1)$. The factor $1/\Gamma(s)$ cancels all these, except the pole at $s = 1$. It follows that $\zeta$ extends meromorphically with a simple pole at $s = 1$.

In view of Exercise 9 the value at $s + 2k - 1 = 0$ is

$$\frac{B_{2k}}{(2k)!} \cdot (-1)^{2k-1}(2k-1)! \;=\; -\frac{B_{2k}}{2k}.$$

Putting these pieces together gives the following.

**Theorem 11.1.1.** *The zeta function is an entire meromorphic function. The only pole is a simple pole at $s = 1$ with residue 1. Moreover,*

$$\zeta(-2n) = 0, \quad n \;=\; 1, 2, 3, \ldots\;;$$
$$\zeta(0) = -\frac{1}{2};$$
$$\zeta(1-2k) = -\frac{B_{2k}}{2k}, \quad k \;=\; 1, 2, 3, \ldots\;.$$

The location of the *other* zeros of $\zeta$ is a matter of some interest; see Chapter 13.

## 11.2 The functional equation of the zeta function

There is another way, due to Riemann, to extend $\zeta$. Consider the integral

$$I_\delta(s) \;=\; \frac{1}{2\pi i} \int_{C_\delta} \frac{(-x)^s}{e^x - 1}\, \frac{dx}{x}. \tag{11.2.1}$$

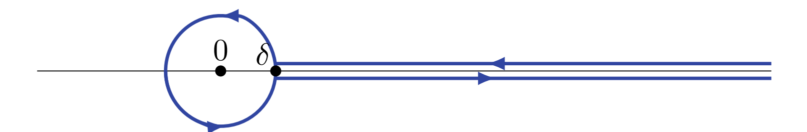



**Fig. 11.1** Riemann's contour $C_\delta$

Here $C_\delta$ is a contour that goes from $+\infty$ to $\delta > 0$ along the positive real axis, goes in the positive direction around the circle $\{z : |z| = \delta\}$, and returns to $+\infty$ along the positive real axis; see Figure 11.1. It is assumed that $\delta < 2\pi$, so the integrand is holomorphic for $0 < |z| \le \delta$.

This takes some interpretation, particularly if $s$ is not an integer. We take the principal branch of the power, with $(-x)^s$ holomorphic for $x$ in $\Omega = \mathbb{C} \setminus [0, \infty)$, the complement of the non-negative real axis.

Along the first part of the contour, the argument of $-x$ is $-i\pi$, so the integral along this part is

$$-e^{-i\pi s} \int_\delta^\infty \frac{x^{s-1}}{e^x - 1}\, dx.$$

Similarly, the integral along the final part of the contour is

$$e^{i\pi s} \int_\delta^\infty \frac{x^{s-1}}{e^x - 1}\, dx,$$

so

$$\begin{aligned} I_\delta(s) &= \frac{e^{i\pi s} - e^{-i\pi s}}{2\pi i} \int_\delta^\infty \frac{x^{s-1}}{e^x - 1}\, dx + (\text{integral around } \{z : |z| = \delta\}) \\ &= \frac{\sin \pi s}{\pi} \int_\delta^\infty \frac{x^s}{e^x - 1} \frac{dx}{x} + (\text{integral around } \{z : |z| = \delta\}). \end{aligned} \tag{11.2.2}$$

Several things should be noted here:

1. $I_\delta(s)$ is independent of $\delta > 0$ (Cauchy's theorem).

2. $I_\delta(s)$ is an entire function of $s$. In fact $z^s$ is an entire function of $s$ for each $z \ne 0$, and the growth of $e^x$ guarantees convergence of the integral.

3. When $\operatorname{Re} s > 1$ the integral over the circle has limit 0 as $\delta \to 0$; Exercise 10. Thus if $\operatorname{Re} s > 1$, we may let $\delta \to 0$ and conclude that

$$I_\delta(s) = \frac{\sin \pi s}{\pi} \int_0^\infty \frac{x^{s-1}}{e^x - 1}\, dx.$$

Taking into account the reflection formula for the gamma function, (10.4.1), together with (11.1.1), we have

$$I_\delta(s) = \frac{1}{\Gamma(s)\Gamma(1-s)} \int_0^\infty \frac{x^{s-1}}{e^x - 1}\, dx$$
$$= \frac{\zeta(s)}{\Gamma(1-s)}, \qquad \mathrm{Re}\, s > 1. \tag{11.2.3}$$

4. Since $I_\delta$ is an entire function of $s$, the identity (11.2.3) extends to all $s \neq 1, 2, 3, \ldots$. In particular, we shall want to use (11.2.3) for $\mathrm{Re}\, s < 0$. This leads to Riemann's integral formula that extends $\zeta$:

$$\zeta(s) \;=\; \frac{\Gamma(1-s)}{2\pi i} \int_{C_\delta} \frac{(-x)^s}{e^x - 1}\, \frac{dx}{x}. \tag{11.2.4}$$

5. The integrand is meromorphic in the domain "bounded" by $C_\delta$, with poles where $e^z = 1$, i.e. $z = 2n\pi i$, $n = \pm 1, \pm 2, \ldots$. The residues at these points are certain multiples of $|n|^{s-1}$. It follows that if the integral around large circles $\{z : |z| = R\}$ tends to zero, then $I_\delta(s)$ for $\mathrm{Re}\,(1-s) > 1$ may be a multiple of $\zeta(1-s)$. If so, then we obtain a relation between $\zeta(s)$ and $\zeta(1-s)$: the functional equation for $\zeta$.

**Theorem 11.2.1.** *The zeta function satisfies the identity*

$$\zeta(s) \;=\; 2\Gamma(1-s)\,(2\pi)^{s-1} \sin\left(\frac{\pi s}{2}\right) \zeta(1-s). \tag{11.2.5}$$

*Proof:* As suggested above, the idea is to compute the integral (11.2.4) by the residue calculus. We begin with a computation of the residues, and then justify the argument. (Note that the negative sign in (11.2.4) corresponds to changing the curve $C_\delta$, so that the following residue argument is applicable.)

The residues occur at the singularities of $1/(e^z - 1)$, which are the points $z = 2n\pi i$, $n \neq 0$. Each is a simple pole $1/(e^z - 1)$ with residue 1. For $n$ a positive integer,

$$(-in)^{s-1} + (in)^{s-1} = \left[e^{-i\frac{1}{2}\pi(s-1)} + e^{i\frac{1}{2}\pi(s-1)}\right] n^{s-1}$$
$$= -i\left[e^{I\frac{1}{2}\pi s} - e^{-i\frac{1}{2}\pi s}\right] n^{s-1} \;=\; 2\sin(\pi s/2) n^{s-1}.$$

Assuming that the residue calculus is legitimate here,

$$\zeta(s) \;=\; \Gamma(1-s) \sum_{n=1}^{\infty} (2n\pi)^{s-1} 2\sin\left(\frac{\pi s}{2}\right).$$

Thus, formally, we have derived (11.2.5).

To justify this calculation, let

$$\Omega_{\delta,n} \;=\; \{z \in \Omega_\delta : |z| < (2n+1)\pi\}.$$

The residue calculus applies to each domain $\Omega_{\delta,n}$, so to complete the proof we need only to show that for $\operatorname{Re} s < 0$,

$$\lim_{n\to\infty} \int_{|z|=(2n+1)\pi} \frac{z^{s-1}}{e^z - 1}\, dz \;=\; 0.$$

The integral of $|z^{s-1}|$ over this circle is $O(n^{\operatorname{Re} s})$, so we can establish (11.2.5) for $\operatorname{Re} s < 0$ by showing that $1/(e^z - 1)$ is bounded on the circle, uniformly with respect to $n$. The validity for every $s$ follows by analytic continuation.

Let $r = (2n+1)\pi$, so the circle is $\{re^{i\theta} : 0 \le \theta < 2\pi\}$. Now

$$\exp(re^{i\theta}) \;=\; e^{r\cos\theta}\, e^{ir\sin\theta}$$

has modulus $e^{r\cos\theta}$, so the distance from $\exp(re^{i\theta})$ to $z = 1$ is at least $1 - 1/e$ unless $-1 \le r\cos\theta \le 1$. In this case $\cos^2\theta \le 1/r^2$, so

$$(1-\sin\theta)(1+\sin\theta) \;=\; 1 - \sin^2\theta \le \frac{1}{r^2}.$$

If $\sin\theta > 0$, we have $1 - \sin\theta \le 1/r^2$. Thus, for each $2m\pi i$,

$$\begin{aligned}|r\sin\theta - 2m\pi| \ge |r - 2m\pi| - |r - r\sin\theta| \;&\ge\; |2n+1-2m|\pi - \frac{1}{(2n+1)\pi}\\ &\ge \pi - \frac{1}{(2n+1)\pi} \;>\; \frac{\pi}{2}.\end{aligned}$$

The same argument applies if $\sin\theta < 0$. This shows that if the modulus of $\exp(re^{i\theta})$ is close to 1, then the argument differs from every $2m\pi$ by more than $\pi/2$. Thus $1/(e^z - 1)$ is bounded on the curve, uniformly with respect to $n$. $\square$

## 11.3 Zeros

The following was proved independently by de la Vallée Poussin and Hadamard, and used in their proofs of the prime number theorem. We follow de la Vallée Poussin's argument here.

**Theorem 11.3.1.** *The zeta function has no zeros on the line* $\{s : \operatorname{Re} s = 1\}$.

*Proof:* Consider the negative of the logarithmic derivative

$$-\frac{\zeta'(s)}{\zeta(s)} \;=\; \sum_p \left[\log(1-p^{-s})\right]' \;=\; \sum_p \frac{\log p \cdot p^{-s}}{1-p^{-s}} \;=\; \sum_p \frac{\log p}{p^s - 1}.$$

We would prefer a simpler denominator, like $p^s$. Since

$$\frac{1}{p^s-1} = \frac{1}{p^s}+\left\{\frac{1}{p^s-1}-\frac{1}{p^s}\right\} = \frac{1}{p^s}+\frac{1}{p^s(p^s-1)},$$

we pass to

$$-\frac{\zeta'(s)}{\zeta(s)} = \sum_p \frac{\log p}{p^s}+\sum_p \frac{\log p}{p^s(p^s-1)} \equiv \Phi(s)+\sum_p \frac{\log p}{p^s(p^s-1)}. \tag{11.3.1}$$

The last sum converges for $\operatorname{Re} s > 1/2$, since it is dominated by $\sum(\log n/n^{2s})$. Therefore $\Phi$ and $-\zeta'/\zeta$ have the same singularities for $\operatorname{Re} s = 1$. In particular, if $\zeta$ has a zero of order $m$ at $1+i\alpha$, $\alpha > 0$, then the residue of $-\zeta'/\zeta$ at this point is

$$\lim_{\varepsilon\to 0} \varepsilon\,\Phi(1+\varepsilon+i\alpha) = -m. \tag{11.3.2}$$

If $\alpha = 0$, this limit is 1.

Suppose that $\beta > 0$ and suppose that $\zeta$ has a zero of order $m$ at $1+2i\beta$ and a zero of order $k$ (possibly zero) at $1+4i\beta$. By the reflection principle, $\zeta$ vanishes to the same orders at $1-2i\beta$ and $1-4i\beta$. Now $k \geq 0$, and we want to show that $m = 0$, i.e. $\zeta$ does not vanish at $1\pm 2i\beta$. For $\varepsilon > 0$,

$$\begin{aligned} 0 &< \varepsilon \sum_p \frac{\log p}{p^{1+\varepsilon}}(p^{i\beta}+p^{-i\beta})^4 \\ &= \varepsilon\,[\Phi(1+\varepsilon-4i\beta)+4\,\Phi(1+\varepsilon-2i\beta)+6\,\Phi(1+\varepsilon) \\ &\qquad +4\,\Phi(1+\varepsilon+2i\beta)+\Phi(1+\varepsilon+4i\beta)]. \end{aligned}$$

Taking the limit as $\varepsilon \downarrow 0$,

$$0 \leq -k-4m+6-4m-k = 6-2k-8m,$$

so $m = 0$. $\square$

## 11.4 $\zeta(2m)$

We used Euler's product formula for sine to evaluate $\zeta(2)$. In principle the product formula can be used to evaluate $\zeta$ at each positive even integer. Euler did this by taking the derivative of the logarithm of both sides, using the right side to find the Maclaurin series and expressing the left side in terms of $e^{i\pi z}$.

Let us carry this out. Taking the derivative of the logarithm of both sides of the product formula

$$\sin \pi z = \pi z \prod_{n=1}^{\infty}\left(1+\frac{z}{n}\right)\left(1-\frac{z}{n}\right)$$

and subtracting $1/z$ from each side give

$$\frac{\pi\cos\pi z}{\sin\pi z} - \frac{1}{z} = \sum_{n=1}^{\infty}\left(\frac{1}{n+z} - \frac{1}{n-z}\right). \tag{11.4.1}$$

Each side is holomorphic for $|z| < 1$. The $k$-th derivative of the right side is

$$\sum_{n=1}^{\infty}\left\{\frac{(-1)^k k!}{(n+z)^{k+1}} - \frac{k!}{(n-z)^{k+1}}\right\}.$$

Thus the Maclaurin expansion (the Taylor expansion at $z = 0$) of the right side of (11.4.1) is

$$-2\sum_{m=1}^{\infty}\zeta(2m)z^{2m-1}. \tag{11.4.2}$$

On the other hand, taking into account (11.1.3), the left side of (11.4.1) is

$$\begin{aligned} i\pi\frac{e^{i\pi z}+e^{-i\pi z}}{e^{i\pi z}-e^{-i\pi z}} - \frac{1}{z} &= i\pi\frac{e^{2i\pi z}+1}{e^{2i\pi z}-1} - \frac{1}{z} \\ &= 2i\pi\left[\frac{1}{e^{2i\pi z}-1} - \frac{1}{2i\pi z} + \frac{1}{2}\right] \\ &= 2i\pi\sum_{m=1}^{\infty}\frac{B_{2m}}{(2m)\,!}(2i\pi z)^{2m-1}. \end{aligned} \tag{11.4.3}$$

Comparing (11.4.2) and (11.4.3) gives

$$\zeta(2m) = \frac{(-1)^{m-1}}{2}\frac{(2\pi)^{2m}}{(2m)\,!}B_{2m}. \tag{11.4.4}$$

We have seen that $\zeta$ vanishes at the negative even integers. The values at the positive odd integers $n$, $n > 1$, are not well understood.

## 11.5 The function $\xi(s)$

Riemann introduced a function $\xi$ that is useful in the study of $\zeta(s)$ (particularly in connection with the Riemann hypothesis). It has three important properties:

(i) $\xi$ is an entire function.

(ii) The zeros of $\xi$ are the zeros of $\zeta$ that lie in the strip $\{0 < \mathrm{Re}\, s < 1\}$ (the "non-trivial zeros" of $\zeta$).

(iii) $\xi$ is symmetric about the line $\{s : \mathrm{Re}\, s = \frac{1}{2}\}$, i.e. $\xi(s) = \xi(1-s)$.

Riemann simply wrote down a formula for $\xi$, but we can guess how he might have reasoned. First, it is easy to accomplish (i) and (ii). Multiplying $\zeta$ by $(s-1)$ gets rid of the pole of $\zeta$, and multiplication by $\Gamma(\frac{1}{2}s+1) = \frac{1}{2}s\Gamma(\frac{1}{2}s)$ turns the zeros $-2, -4, -6, \ldots$ into removable singularities. Therefore our first guess might be

$$\xi_0(s) \;=\; \frac{s(s-1)}{2}\Gamma\left(\frac{s}{2}\right)\zeta(s).$$

Then

$$\xi_0(1-s) \;=\; \frac{s(s-1)}{2}\Gamma\left(\frac{1}{2}-\frac{s}{2}\right)\zeta(1-s).$$

The next step is to look at $\xi_0$ in the context of the functional equation (11.2.5). It is convenient here to use the reflection formula (10.4.1) and write

$$\sin\left(\frac{\pi s}{2}\right) \;=\; \frac{\pi}{\Gamma\left(\frac{s}{2}\right)\Gamma\left(1-\frac{s}{2}\right)},$$

so that (11.2.5) becomes

$$\zeta(s) \;=\; \frac{\Gamma(1-s)(2\pi)^s}{\Gamma\left(\frac{s}{2}\right)\Gamma\left(1-\frac{s}{2}\right)}\zeta(1-s).$$

Then there is a cancellation, and

$$\begin{aligned}\xi_0(s) &= \frac{\Gamma(1-s)(2\pi)^s}{\Gamma\left(1-\frac{s}{2}\right)}\cdot\frac{s(s-1)}{2}\zeta(1-s)\\ &= \frac{\Gamma(1-s)(2\pi)^s}{\Gamma\left(1-\frac{s}{2}\right)\Gamma\left(\frac{1}{2}-\frac{s}{2}\right)}\xi_0(1-s).\end{aligned} \tag{11.5.1}$$

By Legendre's duplication formula (10.3.2),

$$\Gamma(1-s) \;=\; \pi^{-1/2}2^{-s}\Gamma\left(\frac{1}{2}-\frac{s}{2}\right)\Gamma\left(1-\frac{s}{2}\right),$$

so (11.5.1) simplifies to

$$\xi_0(s) \;=\; \pi^{s-\frac{1}{2}}\xi_0(1-s)$$

or, equivalently,

$$\pi^{-s/2}\xi_0(s) \;=\; \pi^{-(1-s)/2}\xi_0(1-s).$$

Thus we take $\xi(s) = \pi^{-s/2}\xi_0(s)$, i.e.

$$\xi(s) \;=\; \pi^{-s/2}\frac{s(s-1)}{2}\Gamma\left(\frac{s}{2}\right)\zeta(s). \tag{11.5.2}$$

**Theorem 11.5.1.** (Riemann) *The function $\xi(s)$ defined by (11.5.2) is entire, and is symmetric about the line $\{s : \operatorname{Re} s = 1/2\}$:*

$$\xi(1-s) \;=\; \xi(s). \tag{11.5.3}$$

*Moreover*

$$\xi(0) \;=\; \xi(1) \;=\; \frac{1}{2}. \tag{11.5.4}$$

*Proof:* The preceding argument established everything except (11.5.4). In view of the symmetry, we need to only evaluate $\xi(0)$. But $(s/2)\Gamma(s/2) \to 1$ as $s \to 0$, so

$$\xi(0) = \Big[(s-1)\pi^{-s/2}\zeta(s)\Big]_{s=0} = -\zeta(0) = \frac{1}{2}. \qquad \square \tag{11.5.5}$$

## Exercises

1. Show that Euler's product (11.0.2) converges when $\mathrm{Re}\, s > 1$.
2. Show that for $\mathrm{Re}\, s > 1$,
$$\frac{1}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\mu(n)}{n^s},$$
where $\mu$ is the *Möbius function*: $\mu(1) = 1,\ \mu(n) = -1$ if $n$ is the product of an odd number of distinct primes, $\mu(n) = 1$ if $n$ is the product of an even number of distinct primes, and $\mu(n) = 0$ if $n$ has a repeated prime factor.
3. Show that for $\mathrm{Re}\, s > 1$,
$$\frac{\zeta(s)}{\zeta(2s)} = \sum_{n=1}^{\infty} \frac{|\mu(n)|}{n^s}.$$
4. Show that for $\mathrm{Re}\, s > 1$,
$$[\zeta(s)]^2 = \sum_{n=1}^{\infty} \frac{d(n)}{n^s},$$
where $d(n)$ is the number of divisors of $n$.
5. The prime number theorem implies that for every $\varepsilon > 0$, there are infinitely many primes greater than $(1-\varepsilon)x/\log x$. Show that the divergence of $\sum 1/p$ already implies the weaker result: for every $\varepsilon > 0$, there are infinitely many primes greater than $x/(\log x)^{1+\varepsilon}$. (Hint: suppose that this expression is bounded, and estimate $\sum 1/p$ for $2^{n-1} < p < 2^n$.)
6. Use the trivial inequality
$$\sum_{k=1}^{n} \frac{1}{k} \le \prod_{p \le n} \left(1 - \frac{1}{p}\right)^{-1}$$
to prove an inequality that is half of (11.0.3).
7. Use the (non-trivial) fact that $\sum_{p \le n} \log p \sim n$ as $n \to \infty$ (see Lemma 13.3.3) to prove the other half of (11.0.3).
8. Show that for $\mathrm{Re}\, s > 1$,
$$\log \zeta(s) = \sum_{p} \sum_{n=1}^{\infty} \frac{1}{n\, p^{ns}}.$$
9. Compute $\zeta'(s)/\zeta(s)$ for $\mathrm{Re}\, s > 0$.
10. Verify that for $\mathrm{Re}\, s > 1$, the portion of the integral (11.2.1) over the circle $\{z : |z| = \delta\}$ is $O(\delta)$ as $\delta \to 0$.

11. Compute $\xi'(s)/\xi(s)$ for $\operatorname{Re} s > 1$.
12. Show that all the zeros of $\xi(s)$ lie in the open strip $\{s : 0 < \operatorname{Re} s < 1\}$.
13. It can be shown that $\xi$ has a factorization

$$\xi(s) = \xi(0)\prod_{\rho}\left(1 - \frac{s}{\rho}\right),$$

where the $\rho$ are the zeros of $\xi$, repeated according to multiplicity and the product is taken to be

$$\lim_{R\to\infty}\prod_{|\rho|\le R}\left(1 - \frac{s}{\rho}\right) = \prod_{\rho}\left(1 - \frac{s(1-s)}{\rho(1-\rho)}\right);$$

see Section 8.5. Find a formula for $\xi'(s)/\xi(s)$ that is valid for every $s$ for which $\xi(s) \neq 0$.

14. For $\operatorname{Re} s > 1$, let

$$L(s) = 1 - \frac{1}{3^s} + \frac{1}{5^s} - \frac{1}{7^s} + \dots.$$

Show that $L$ extends to an entire function.

15. Show that the function $L$ of the previous problem satisfies the functional equation

$$L(1-s) = \left(\frac{2}{\pi}\right)^s \pi^{-s}\sin\left(\frac{\pi s}{2}\right)\Gamma(s)L(s).$$

## Remarks and further reading

For much more about the zeta function, see Edwards [42], Ivic [72], and Titchmarsh [135]. See also Chapter 13 and the references there.

# Chapter 12
# L-functions and primes




Euler used his product expansion of the Riemann zeta function

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right)^{-1}, \qquad \operatorname{Re} s > 1$$

to give an analytic proof that there are infinitely many primes. Dirichlet extended this result to prove that there are infinitely many primes in arithmetic progressions.

**Theorem 12.0.1.** (Dirichlet) *If $k$ and $m$ are relatively prime integers (i.e. have no common factors), $0 < m < k$, then there are infinitely many primes in the arithmetic progression*

$$m,\ k+m,\ 2k+m,\ 3k+m,\ \ldots\ . \tag{12.0.1}$$

Note that if a prime $p$ does not divide $k$, then dividing $k$ into $p$ leaves a remainder $0 < m < k$ that is relatively prime to $k$. It follows that for *some* remainder $m$, there are infinitely many primes $p$ in the progression (12.0.1). Dirichlet's theorem states that this is true not just for one such remainder $m$, but for *each* relatively prime remainder mod $k$. For example, there are infinitely many primes of the form $4n+1$ and also infinitely many of the form $4n+3$.

This chapter is devoted principally to the proof of Theorem 12.0.1. On the way, we develop the theory of characters for a finite commutative group. In the last section and the exercises we take up the question of functional equations for (certain) $L$ functions, analogous to the functional equation for the zeta function.

A comment on the notation here, in which $k$ is taken to be some fixed number and $m$ is an index, $1 \le m < k$. Awkward and unusual as this may seem, it is the standard convention for this topic.



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_12

## 12.1 Factorization and Dirichlet characters

Dirichlet's fundamental observation was that Euler's product formula could be extended to series of the form

$$L(s,\chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} \tag{12.1.1}$$

provided that the function $\chi : \mathbb{N} \to \mathbb{C}$ is bounded and *multiplicative*:

$$\chi(mn) = \chi(m)\,\chi(n), \qquad m,n = 1,2,3,\dots . \tag{12.1.2}$$

These assumptions imply that $\chi$ takes values in $\{0\} \cup \{z : |z| = 1\}$. By comparison with $\zeta$, the series $L(\cdot,\chi)$ converges for $\operatorname{Re} s > 1$. Moreover, it follows from (12.1.2) and unique factorization of integers that $L(\cdot,\chi)$ satisfies the analogue of Euler's product formula

$$L(s,\chi) = \prod_{p \text{ prime}} \left(1 - \frac{\chi(p)}{p^s}\right)^{-1}. \tag{12.1.3}$$

The functions $\chi$ used in the proof of Theorem 12.0.1 depend on the choice of the integer $k$. For convenience we use some standard notation from number theory: if $k$ and $m$ are positive integers, $(m,k)$ denotes the largest common factor of $k$ and $m$. Thus $(m,k) = 1$ means that there are no non-trivial common factors: $m$ and $k$ are relatively prime. Integers $m,n$ are said to be equal mod $k$, written $m = n$ mod $k$, if $k$ divides $m - n$.

A *Dirichlet character mod k* is a map

$$\chi : \mathbb{N} \to \{0\} \cup \{z : |z| = 1\} \tag{12.1.4}$$

that is multiplicative and has the properties:

$$\chi(n) = \chi(m) \text{ if } n = m \mod k; \tag{12.1.5}$$

$$\chi(n) \neq 0 \text{ if and only if } (n,k) = 1. \tag{12.1.6}$$

Thus a Dirichlet character is periodic with period $k$. Moreover it is uniquely determined by its values on the set of remainders mod $k$:

$$G_k = \{m : (m,k) = 1,\ 1 \le m < k\}.$$

The set $G_k$ is a commutative group with respect to the operation of multiplication mod $k$. (The existence of an inverse is a counting argument: for fixed $m \in G_k$, the remainders of $mn$ are distinct as $n$ runs through $G_k$, so one of these remainders is 1.) The restriction to $G_k$ of a Dirichlet character $\chi$ mod $k$ is a *group character* in the standard sense:

$$\chi : G_k \to \{z : |z| = 1\}, \quad \chi(mn) = \chi(m)\,\chi(n).$$

Conversely, a group character on $G_k$ extends uniquely to a Dirichlet character mod $k$ by using (12.1.5) and (12.1.6) to determine the extension.

## 12.2 Characters of finite commutative groups

We need some algebraic information that uses no properties that are unique to the group $G_k$, so we develop the character theory of an arbitrary finite commutative group $G$. We write the composition in $G$ multiplicatively, and denote the identity element by 1. We denote by $\widehat{G}$ the set of characters of $G$. Let $|G|$ denote the *order* of $G$, i.e. the number of elements in $G$. We assume $|G| > 1$.

**Lemma 12.2.1.** *If $\chi_1$ and $\chi_2$ are distinct characters of $G$, then*

$$\sum_{g\in G} \chi_1(g)\overline{\chi_2}(g) \;=\; 0. \tag{12.2.1}$$

*Proof:* Note that $\chi = \chi_1\overline{\chi_2}$ is itself a character. For each $h$ in $G$,

$$\left[\sum_{g\in G} \chi(g)\right] \chi(h) \;=\; \sum_{g\in G} \chi(gh) \;=\; \sum_{g\in G} \chi(g), \tag{12.2.2}$$

since $gh$ runs through $G$ as $g$ runs through $G$. The identity (12.2.2) implies that either the sum is zero or $\chi(h) = 1$ for each $h$:

$$\sum_{g\in G} \chi(g) \;=\; \begin{cases} |G| & \text{if } \chi \equiv 1; \\ 0 & \text{if } \chi \not\equiv 1. \end{cases} \tag{12.2.3}$$

If $\chi_1 \neq \chi_2$, then $\chi$ is not identically 1, so (12.2.1) follows from (12.2.3). □

**Lemma 12.2.2.** *Given $g \neq 1$ in $G$, there is a character $\chi$ such that $\chi(g) \neq 1$.*

*Proof:* The set of elements $1, g, g^2, \ldots$ is a subgroup $H_g$ of $G$. There is a smallest positive integer $m$ such that $g^m = 1$. Let $\omega = e^{2\pi i/m}$, and let $\chi(g^k) = \omega^k$. Then $\chi$ is a character of $H_g$ and $\chi(g) \neq 1$. Suppose that $\chi$ has been extended to a subgroup $H$ of $G$. If $H \neq G$, we shall show that $\chi$ can be extended to a larger subgroup; thus, in finitely many steps $\chi$ can be extended to all of $G$. Suppose that $g_1$ is not in $H$. There is some smallest power $n$ such that $g_1^n$ is in $H$. Choose $\omega_1$ such that $(\omega_1)^n = \chi(g_1^n)$ and define

$$\chi(g_1^m h) \;=\; \omega_1^m \chi(h), \quad h \in H, \quad 1 \le m < n.$$

Then

$$\chi(g_1^m h)\,\chi(g_1^{m'} h') \;=\; (\omega_1)^{m+m'} \chi(hh')$$

and it is easily checked in the two cases $m + m' < n$ and $m + m' \geq n$ that the right side is $\chi(g_1^m h g_1^{m'} h')$. □

The set of characters $\widehat{G}$ is itself a group under the usual operation of multiplication of functions:

$$[\chi_1\chi_2](gh) = \chi_1(gh)\chi_2(gh) = \ldots = [\chi_1\chi_2](g)\cdot[\chi_1\chi_2](h).$$

We can now identify certain characters of the character group $\widehat{G}$. Given $g \in G$, define

$$\xi_g(\chi) = \chi(g), \quad \chi \in \widehat{G}.$$

This is a homomorphism (i.e. products go to products): check that $\xi_{gh} = \xi_g\xi_h$.

**Lemma 12.2.3.** *The map $g \to \xi_g$ from $G$ to the character group of $\widehat{G}$ is injective. In particular*

$$|G| \leq |(\widehat{G})\widehat{\ }|. \tag{12.2.4}$$

*Proof:* Suppose that $g$ and $h$ are distinct elements of $G$. Then $gh^{-1} \neq 1$, and according to Lemma 12.2.2 there is some $\chi \in \widehat{G}$ such that

$$1 \neq \chi(gh^{-1}) = \chi(g)\,\chi(h^{-1}) = \xi_g(\chi)\,[\xi_h(\chi)]^{-1},$$

so $\xi_g \neq \xi_h$. $\square$

The next, and very important, step is to construct a certain finite dimensional inner product space of functions defined on $G$:

$$\mathscr{F} = \{u : \ u : G \to \mathbb{C}\}.$$

This is a complex vector space of dimension $|G|$. We take as inner product

$$(u,w) = \frac{1}{|G|}\sum_{g\in G} u(g)\,\overline{w(g)}.$$

**Proposition 12.2.4.** *The set of characters $\widehat{G}$ is an orthonormal basis for $\mathscr{F}$.*

*Proof:* It follows from Lemma 12.2.1 that the $\widehat{G}$ is an orthonormal set in $\mathscr{F}$. Therefore the characters are linearly independent. It follows from this fact that

$$|\widehat{G}| \leq |G|.$$

This is a general inequality for finite commutative groups, so we may apply it to $\widehat{G}$ and then use (12.2.4) to conclude that

$$|G| \leq |(\widehat{G})\widehat{\ }| \leq |\widehat{G}|.$$

Thus $|G| = |\widehat{G}|$, so $|\widehat{G}|$ equals the dimension of $\mathscr{F}$. Thus $\widehat{G}$ is a basis for $\mathscr{F}$. $\square$

As we have just noted, $|G| = |\widehat{G}|$, so the map $g \to \xi_g$ from $G$ to the character group of $\widehat{G}$ is a surjective isomorphism.

**Lemma 12.2.5.** *If $g \in G$, then*

$$\sum_{\chi \in \widehat{G}} \chi(g) = \begin{cases} 0 & \text{if } g \neq 1; \\ |\widehat{G}| & \text{if } g = 1. \end{cases} \tag{12.2.5}$$

*Proof:* In light of the preceding remarks, this is (12.2.3), with $G$ replaced by $\widehat{G}$ and $g$ replaced by $\xi_g$ in the character group of $\widehat{G}$:

$$\sum_{\chi \in \widehat{G}} \chi(g) = \sum_{\chi \in \widehat{G}} \xi_g(\chi). \qquad \square$$

## 12.3 Analysis of $L$-functions

We now turn to the analytical part of the proof of Theorem 12.0.1.

The fact that a Dirichlet character mod $k$ is periodic mod $k$ allows us to analyze the possibility of analytic extension.

**Proposition 12.3.1.** *Suppose that $\chi : \mathbb{N} \to \mathbb{C}$ is periodic with period $k$, and let $L(s,\chi)$ be defined by*

$$L(s,\chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s}, \quad \operatorname{Re} s > 1.$$

*Then $L(\cdot,\chi)$ has an extension that is holomorphic in the plane, with the possible exception of a simple pole at $s = 1$ with residue*

$$\sum_{m=1}^{k} \frac{\chi(m)}{k}. \tag{12.3.1}$$

*Proof:* Each positive integer $N$ can be written uniquely as $nk+m$ for some $1 \leq m \leq k$, $n \geq 0$. Since $\chi(N) = \chi(m)$, we may regroup the series that defines $L(s,\chi)$:

$$L(s,\chi) = \sum_{j=1}^{k} \chi(j)\phi_{j,k}(s), \tag{12.3.2}$$

where

$$\phi_{j,k}(s) = \sum_{n=0}^{\infty} \frac{1}{(j+nk)^s}. \tag{12.3.3}$$

As with the earlier discussion of $\zeta$, we use the identity

$$\frac{1}{a^s} = \frac{1}{\Gamma(s)} \int_0^{\infty} e^{-at} t^{s-1} dt, \qquad \operatorname{Re} a > 0, \ \operatorname{Re} s > 1,$$

together with uniform convergence, to write

$$\begin{aligned}\phi_{j,k}(s) &= \frac{1}{\Gamma(s)} \sum_{n=0}^{\infty} \int_0^{\infty} e^{-(j+nk)t} t^{s-1} dt \\ &= \frac{1}{\Gamma(s)} \int_0^{\infty} e^{-jt} \sum_{n=0}^{\infty} (e^{-kt})^n t^{s-1} dt \\ &= \frac{1}{\Gamma(s)} \int_0^{\infty} \frac{e^{-jt}}{1-e^{-kt}} t^{s-1} dt \\ &= \frac{1}{\Gamma(s)} \left[ \int_0^{\varepsilon} \frac{e^{-jt}}{1-e^{-kt}} t^{s-1} dt + \int_{\varepsilon}^{\infty} \frac{e^{-jt}}{1-e^{-kt}} t^{s-1} dt \right].\end{aligned} \tag{12.3.4}$$

For each $\varepsilon > 0$, the second integral in the last line extends to an entire function of $s$. The first integrand, as a function of $t$, is holomorphic for $|kt| < 2\pi$, except for a simple pole at $t = 0$ with residue $1/k$. Therefore, for each positive $\varepsilon < 2\pi/k$, and $\operatorname{Re} s > 1$, the first integral has the form

$$\int_0^{\varepsilon} \left\{ \frac{1}{kt} + \sum_{n=0}^{\infty} a_n^{jk} t^n \right\} t^{s-1} dt,$$

and the series is uniformly convergent on the interval $[0, \varepsilon]$. Thus the first integral is

$$\frac{\varepsilon^{s-1}}{k(s-1)} + \sum_{n=0}^{\infty} \frac{a_n^{jk} \varepsilon^{n+s}}{n+s}.$$

This defines a function of $s$ that is meromorphic in the plane. It has a simple pole at $s = 1$ with residue $1/k$. Poles at the non-positive integers are killed by the multiplication by $1/\Gamma(s)$, which vanishes at the non-positive integers. □

We now return to $G_k$, with the group characters extended to be Dirichlet characters mod $k$ on $\mathbb{N}$. Let $\chi_1$ denote the *principal character*: $\chi_1 \equiv 1$. Thus

$$L(s, \chi_1) = \sum_{(n,k)=1} \frac{1}{n^s}.$$

Proposition 12.3.1 and equation (12.2.3) applied to the $L(\cdot, \chi)$ give the following.

**Proposition 12.3.2.** *The function $L(s, \chi)$ extends to an entire function of $s$ if $\chi \neq \chi_1$. The function $L(s, \chi_1)$ extends to a function with one pole, a simple pole at $s = 1$ with residue $|G_k|/k$.*

**Corollary 12.3.3.** *(a) The limit*

$$\lim_{s \to 1+} \sum_p \frac{\chi(p)}{p^s} = L(1, \chi) \tag{12.3.5}$$

*is $+\infty$ if $\chi = \chi_1$ is the principal character mod $m$.*

*(b) If $\chi \neq \chi_1$, then the limit is finite if and only if $L(1, \chi) \neq 0$.*

*Proof:* We know that

$$L(s,\chi) \;=\; \prod_p \left(1-\frac{\chi(p)}{p^s}\right)^{-1}.$$

Taking the principal branch of the logarithm,

$$\log[L(s,\chi)] \;=\; -\sum_p \log\left(1-\frac{\chi(p)}{p^s}\right) \;=\; \sum_p \frac{\chi(p)}{p^s} + R(\chi,s),$$

where $|R(\chi,s)| < \sum(1/n^2) \;=\; \pi^2/6$, for every $\mathrm{Re}\, s > 1$; see (1.5.2). For the principal character $\chi_1$, we know that $L(s,\chi_1)$ blows up as $s \to 1$, so the logarithm blows up, and we have

$$\sum_{(p,k)=1} \frac{1}{p} \;=\; \infty.$$

(This is not surprising, since this sum differs from the sum over *all* primes by finitely many terms, the reciprocals of the prime divisors of $k$.)

Suppose that $\chi$ is not the principal character. Then we know that $L(s,\chi)$ has a finite value at $s=1$. If this value is not zero, then the logarithm is finite at $s=1$ and the previous argument shows that $\sum \chi(p)/p$ is finite. □

## 12.4 Proof of Dirichlet's Theorem

Part (b) of Corollary 12.3.3 shows the importance of the following.

**Theorem 12.4.1.** *If $\chi$ is not the principal character $\chi_1$, then $L(1,\chi) \neq 0$.*

*Proof.* The first step is to show that there is at most one character $\chi$ such that $L(1,\chi)=0$. Consider the product

$$P(s) \;=\; \prod_{\chi \in \widehat{G}_k} L(s,\chi).$$

If more than one character vanishes at $s=1$, then $P(1)=0$. In view of (12.2.5), the logarithm is

$$\sum_p \left[\sum_\chi \chi(p)\right] \frac{1}{p^s} + O(1) \;=\; |G_k| \left[\sum_{p\equiv 1 \bmod k} \frac{1}{p^s}\right] + O(1).$$

This either diverges to $+\infty$ or remains bounded as $s \to 1$, so $P(1) \neq 0$. This proves that at most one $L(s,\chi)$ vanishes at $s=1$.

If $\chi$ is a character, so is its complex conjugate, and the corresponding $L$ functions are conjugates of each other. Since at most one character vanishes at 1, $L(1,\chi) \neq 0$ if $\chi$ is complex, i.e. takes some complex values.

Suppose then that $\chi$ is a real character. Then $\chi$ takes only values $\pm 1$. We follow an argument of de la Vallée Poussin, who introduced the function

$$\begin{aligned}\psi(s) &= \frac{L(s,\chi)L(s,\chi_1)}{L(2s,\chi_1)} \\ &= \prod_{(p,k)=1} \frac{(1-p^{-2s})}{(1-\chi(p)p^{-s})(1-p^{-s})} = \prod_{(p,k)=1} \frac{1+p^{-s}}{1-\chi(p)p^{-s}}.\end{aligned}$$

Factors with $\chi(p) = -1$ drop out, so

$$\begin{aligned}\psi(s) &= \prod_{\chi(p)=1} \frac{1+p^{-s}}{1-p^{-s}} \prod_{\chi(p)=0} (1+p^{-s}) \\ &= \prod_{\chi(p)=1} \left(1+2\sum_{n=1}^{\infty} p^{-sn}\right) \prod_{\chi(p)=0} (1+p^{-s}) \\ &= \sum_{n=1}^{\infty} \frac{a_n}{n^s}, \quad a_n \geq 0.\end{aligned}$$

The derivatives $\psi^{(k)}$ alternate sign:

$$\psi^{(k)}(s) = \sum_{n=1}^{\infty} (-\log n)^k \frac{a_n}{n^s}, \quad \mathrm{Re}\, s > 1. \tag{12.4.1}$$

Suppose that $L(1,\chi) = 0$. This kills the pole of $L(s,\chi_1)$ at $s = 1$. The product representation shows that $L(2s,\chi_1)$ has no zeros in the half plane $\mathrm{Re}\, s > 1/2$. Therefore $\psi$ is regular in the open disk centered at $s = 2$ with radius $3/2$. Moreover $L(2s,\chi_1)$ has a pole at $s = -1/2$, so $\psi$ is regular at $s = 1/2$ and $\psi(1/2) = 0$. Therefore $\psi$ is holomorphic in a slightly larger disk centered at $s = 2$ that contains $s = 1/2$. Then

$$\psi\left(\frac{1}{2}\right) = \sum_{n=0}^{\infty} \frac{\psi^{(n)}(2)}{n\,!} \left(\frac{1}{2} - 2\right)^n. \tag{12.4.2}$$

As noted above, the derivatives $\psi^{(n)}(2)$ alternate sign. Therefore the terms in the series (12.4.2) are non-negative (and not all zero), so $\psi(1/2) > 0$. We have shown that the assumption $L(1,\chi) = 0$ leads to a contradiction. □

It follows from Corollary (12.3.3) and Theorem 12.4.1 that

$$\Sigma_p \frac{\chi(p)}{p} \text{ is finite if } \chi \in \widehat{G}_k,\ \chi \neq \chi_1; \tag{12.4.3}$$

$$\Sigma_p \frac{\chi_1(p)}{p} = +\infty. \tag{12.4.4}$$

We are finally in a position to prove Theorem 12.0.1, in the stronger form given by Dirichlet.

**Theorem 12.4.2.** (Dirichlet) *If $k$ and $m$ are relatively prime integers, $0 < m < k$, then*

$$\sum_{p \equiv m \bmod k} \frac{1}{p} = \infty. \tag{12.4.5}$$

*Proof:* We write (12.4.5) in the equivalent form

$$\sum_p \frac{1_m(p)}{p} = \infty,$$

where the function $1_m : \mathbb{Z} \to \mathbb{R}$ is defined by

$$1_m(n) = \begin{cases} 1 & \text{if } n = m, \\ 0 & \text{if } n \neq m, \end{cases}$$

According to Proposition 12.2.4, we can expand the restriction of $1_m$ to $G_k$ as a linear combination of characters. This expansion extends, by periodicity, to $\mathbb{Z}$, so

$$1_m = \sum_{\chi \in \widehat{G}_k} a(\chi)\,\chi, \qquad a(\chi) = (1_m, \chi).$$

Note that

$$a(\chi_1) = \frac{1}{|G_k|}\chi_1(m) = \frac{1}{|G_k|} > 0. \tag{12.4.6}$$

Then for $\operatorname{Re} s > 1$,

$$\begin{aligned} \sum_p \frac{1_m(p)}{p^s} &= \sum_p \sum_{\chi \in \widehat{G}_k} \frac{a(\chi)\chi(p)}{p^s} \\ &= \sum_{\chi \in \widehat{G}_k} a(\chi) \left[ \sum_p \frac{\chi(p)}{p^s} \right]. \end{aligned}$$

The desired result (12.4.5) follows from (12.4.3), (12.4.4), and (12.4.6). □

## 12.5 Functional equations

Like the zeta function, (certain) Dirichlet $L$-functions satisfy a functional equation that relates $L(s,\chi)$ and $L(1-s,\chi)$. In this section we outline a proof of this statement. Many of the details of the argument are left to the exercises.

Fix $k$. A Dirichlet character $\chi$ mod $k$ is periodic with period $k$ on the positive integers. It can be extended, uniquely, to the negative integers so as to be periodic on $\mathbb{Z}$: $\chi(-n) = \chi(-n+km)$ for $km > n$. Note that a function defined on the integers and having period $k$ can be considered as a function on $\mathbb{Z}/(k)$, the additive group of integers mod $k$. We know from the discussion in Section 12.2 that the characters of

this group are a basis for such functions, and that they are orthonormal with respect to the inner product

$$\langle f,g\rangle = \frac{1}{k}\sum_{n=0}^{k-1} f(n)\overline{g(n)}.$$

The characters $\{\omega_m\}_{m=1}^k$ can be defined in terms of $\alpha = \exp(2\pi i/k)$ by

$$\omega_m(n) = [\omega_1(n)]^m = \alpha^{nm}.$$

Indeed

$$\omega_m(n_1+n_2) = \alpha^{m(n_1+n_2)} = \alpha^{mn_1}\alpha^{mn_2} = \omega_m(n_1)\,\omega_m(n_2),$$

so $\omega_m$ is a character of $\mathbb{Z}/(k)$. It is easily seen that $\langle\omega_m,\omega_m\rangle = 1$, while if $m_1 \neq m_2$ mod $k$ then $\alpha^{m_1-m_2} \neq 1$ and

$$\begin{aligned}\langle\omega_{m_1},\omega_{m_2}\rangle &= \frac{1}{k}\sum_{n=0}^{k-1}\alpha^{(m_1-m_2)n}\\ &= \frac{1}{k}\times\frac{\alpha^{(m_1-m_2)k}-1}{\alpha^{(m_1-m_2)}-1} = 0.\end{aligned}$$

It follows that a Dirichlet character mod $k$ can be written as

$$\chi = \sum_{m=1}^{k}\langle\chi,\omega_m\rangle\,\omega_m, \tag{12.5.1}$$

where

$$\begin{aligned}\langle\chi,\omega_m\rangle &= \frac{1}{k}\sum_{n=1}^{k}\chi(n)\overline{\omega_m(n)}\\ &= \frac{1}{k}\sum_{n=1}^{k}\chi(n)\exp(-2\pi inm/k)\\ &= \frac{1}{k}G(-m,\chi).\end{aligned}$$

In general the *Gauss sum* $G(z,\chi)$ is defined by

$$G(z,\chi) = \sum_{n=1}^{k}\chi(n)\,e^{2\pi izn/k} = \sum_{n=1}^{k}\chi(n)\alpha^{zn}.$$

An important consideration here is that of an *induced modulus*. Suppose that $\widetilde{k}$ is a proper divisor of $k$ and suppose that $\widetilde{\chi}$ is a Dirichlet character mod $\widetilde{k}$. Let $\chi_1$ be the principal character mod $k$. Then $\chi = \widetilde{\chi}\,\chi_1$ is a Dirichlet character mod $k$. In fact $\chi$ is clearly multiplicative, and $(n,k) = 1$ implies $(n,\widetilde{k}) = 1$ and $\chi_1(n) = 1$, so $\chi \neq 0$ if and only if $(n,k) = 1$.

The product (12.1.3) associated to a character mod $k$ effectively only involves primes that do not divide $k$. Therefore if $\widetilde{k}$ is an induced modulus, $\widetilde{\chi}$ is a Dirichlet character mod $\widetilde{k}$, and $\chi$ is the corresponding character mod $k$, then the products (12.1.3) for $L(s,\chi)$ and $L(s,\widetilde{\chi})$ differ only by primes that divide $k$ but not $\widetilde{k}$:

$$L(s,\chi) \;=\; L(s,\widetilde{\chi}) \prod_{p|k,\,(p,\widetilde{k})=1} \left(1-\frac{1}{p^s}\right)^{-1}.$$

The product on the right is finite, so all information about the analytic properties of $L(\cdot,\chi)$ is contained in $L(\cdot,\widetilde{\chi})$.

A character $\chi$ mod $k$ is said to be *primitive* if it has no induced modulus $\widetilde{k} < k$. An $L$ function for a primitive character satisfies a functional equation that is analogous to the functional equation for the zeta function. We need some properties of primitive characters, expressed in terms of the associated Gauss sums $G(\cdot,\chi)$.

**Lemma 12.5.1.** *If $\chi$ is a Dirichlet character mod k and $(n,k)=1$, then*

$$G(n,\chi) \;=\; \overline{\chi}(n)\,G(1,\chi). \tag{12.5.2}$$

*Proof:* If $(n,k)=1$, then $nm$ runs through all residues mod $k$ as $m$ runs from 1 to $k$. Moreover $\chi(n)\overline{\chi}(n)=1$. Therefore

$$\begin{aligned} G(n,\chi) &= \sum_{n=1}^{k} \chi(m)[\chi(n)\overline{\chi}(n)]\alpha^{nm} \\ &= \overline{\chi}(n)\sum_{m=1}^{k} \chi(nm)\alpha^{nm} \\ &= \overline{\chi}(n)\,G(1,\chi). \qquad \square \end{aligned}$$

**Proposition 12.5.2.** *Suppose that $\chi$ is a Dirichlet character mod k, and suppose that there is an n such that $(n,k)=q>0$ and $G(n,\chi)\neq 0$. Then $d=k/q$ is an induced modulus.*

*Proof:* Suppose that $(a,k)=1$, and suppose that $a=1$ mod $d$: $a=1+rd$. In the sum that defines $G(n,\chi)$ we may replace $m$ by $am$:

$$G(n,\chi) \;=\; \sum_{m=1}^{k} \chi(am)\alpha^{nam} \;=\; \chi(a)\sum_{m=1}^{k} \chi(m)\alpha^{nam}. \tag{12.5.3}$$

But

$$\frac{nam}{k} \;=\; \frac{nm+nmrd}{k} \;=\; \frac{nm}{k}+\frac{nmrd}{qd} \;=\; \frac{nm}{k}+\text{integer},$$

since $q$ divides $n$. Therefore $\alpha^{nam}=\alpha^{nm}$ and the identity (12.5.3) reduces to $G(n,\chi)=\chi(a)G(n,\chi)$. By assumption, $G(n,\chi)\neq 0$, so $\chi(a)=1$.

Suppose now that $a$ and $b$ belong to $G_k$ and are equivalent mod $d$. Then there is an element $a' \in G_k$ such that $a'a = 1 \bmod k$. Since $d$ divides $k$, we also have $aa' = 1 \bmod d$, and therefore $ba' = 1 \bmod d$. Consequently

$$\chi(a)\chi(a') = \chi(aa') = 1 = \chi(ba') = \chi(b)\chi(a'),$$

so $\chi(a) = \chi(b)$.

Suppose that every prime that divides $d$ also divides $k$. If $(m,d)=1$ then $(m,k)=1$, and we have just shown that $\chi(m)$ depends only on the equivalence class of $m$ (mod $d$). Thus setting $\widetilde{\chi}(m) = \chi(m)$, $\widetilde{\chi}$ is a character mod $d$.

Suppose finally that $q$ is the product of the primes that divides $k$ but not $d$, and suppose that $(m,d) = 1$. Let $m' = m + qd$. Then $m' = m \bmod d$ and it is easily checked that $(m',k) = 1$. Setting $\widetilde{\chi}(m) = \chi(m')$, we again have a character mod $d$. Thus in either case, $d$ is an induced modulus. □

**Corollary 12.5.3.** *If $\chi$ is a primitive character mod $k$, then for each $n$,*

$$G(n,\chi) = \overline{\chi}(n)\,G(1,\chi). \tag{12.5.4}$$

*Proof:* If $(n,k) = 1$ then Lemma 12.5.1 applies. Otherwise, by Proposition 12.5.2, $G(n,\chi) = 0 = \overline{\chi}(n)$. □

**Proposition 12.5.4.** *Suppose that $\chi$ is a primitive character mod $k$. Then*

$$G(1,\chi)\,G(1,\overline{\chi}) = k\chi(-1). \tag{12.5.5}$$

*Proof:* By Proposition 12.5.2,

$$\begin{aligned} G(1,\chi)\,G(1,\overline{\chi}) &= \sum_{m=1}^{k} G(1,\chi)\overline{\chi}(m)\alpha^m \\ &= \sum_{m=1}^{k} G(m,\chi)\alpha^m \\ &= \sum_{m,n=1}^{k} \chi(n)\alpha^{m+mn}. \end{aligned} \tag{12.5.6}$$

Now

$$\sum_{m=1}^{k} \alpha^{m+mn} = \sum_{m=1}^{k} (\alpha^{n+1})^m.$$

The sum of this geometric series is 0 if $\alpha^{n+1} \neq 1$ and is $k$ if $\alpha^{n+1} = 1$. Thus the sum (12.5.6) collapses to the right side of (12.5.5). □

**Theorem 12.5.5.** *If $\chi$ is a primitive Dirichlet character mod $k$, then the $L$ function $L(s,\chi)$ satisfies the functional equation*

$$L(1-s,\chi) \;=\; \frac{1}{k}\left(\frac{k}{2\pi}\right)^{s} G(1,\chi)\,\Gamma(s)\,L(s,\overline{\chi})\left\{e^{-\pi is/2}+\chi(-1)e^{\pi is/2}\right\}. \quad (12.5.7)$$

For a different version of the functional equation (12.5.7); see Exercise 10.

The rest of this section presents an outline of the proof of Theorem 12.5.5 due to Berndt [20]. It follows from (12.3.2) and (12.3.4) that

$$\Gamma(s)\,L(s,\chi) \;=\; \int_0^\infty \frac{G(ikx/2\pi,\chi)\,x^{s-1}}{1-e^{-kx}}\,dx; \quad (12.5.8)$$

see Exercise 11.

Suppose that $s>1$ and $m$ is a positive integer. Let $C_m$ denote the contour consisting of the semicircle $\{z : |z| = m+\frac{1}{2}, \operatorname{Re} z > 0\}$, together with the imaginary axis from $-i(m+\frac{1}{2})$ to $i(m+\frac{1}{2})$, indented by the semicircle $\{z : |z| = \varepsilon, \operatorname{Re} z > 0\}$, where $\varepsilon < 1$. Let

$$F(z) \;=\; \pi\, e^{-\pi iz}\,\frac{G(z,\overline{\chi})}{G(1,\overline{\chi})}\,\frac{1}{z^s \sin\pi z}.$$

Then $F$ has simple poles at the integers $1,2,\dots,m$, and

$$\frac{1}{2\pi i}\int_{C_m} F(z)\,dz \;=\; \sum_{n=1}^{m} \chi(n)n^{-s}; \quad (12.5.9)$$

see Exercise 11.

There is a uniform bound for $z\in C_m$:

$$\left|\frac{e^{-\pi iz}G(z,\overline{\chi})}{\sin\pi z}\right| \;\le\; M, \quad \text{all } m; \quad (12.5.10)$$

see Exercise 11. It follows that, for $\operatorname{Re} s > 0$, the integral of $F$ over $C_m$ tends to zero as $m\to\infty$, so

$$\begin{aligned} L(s,\chi) \;=\; &\int_{i\varepsilon}^{i\infty} \frac{G(z,\overline{\chi})\,dz}{G(1,\overline{\chi})\,z^s(1-e^{2\pi iz})} + \int_{-i\varepsilon}^{-i\infty} \frac{G(z,\overline{\chi})\,dz}{G(1,\overline{\chi})\,z^s(1-e^{-2\pi iz})} \\ &+\frac{1}{2\pi i}\int_{\Gamma_\varepsilon} F(z)\,dz; \end{aligned} \quad (12.5.11)$$

see Exercise 11. The first two integrals on the right in (12.5.11) continue analytically to all $s\in\mathbb{C}$, so (12.5.11) is valid for every $s$.

Since $\chi(0)=0$, the third integral on the right in (12.5.11) has limit zero as $\varepsilon\to 0$. Therefore for $s<0$,

$$L(s,\chi) = ie^{-\pi is/2}\int_0^\infty \frac{G(iy,\overline{\chi})\,dy}{G(1,\overline{\chi})\,y^s\,(1-e^{-2\pi y})} - ie^{\pi is/2}\int_0^\infty \frac{e^{-2\pi y}\,G(-iy,\overline{\chi})\,dy}{G(1,\overline{\chi})\,y^s\,(1-e^{-2\pi y})}$$
$$= ie^{-\pi is/2}\left(\frac{k}{2\pi}\right)^{1-s}\int_0^\infty \frac{G(iky/2\pi,\overline{\chi})\,dy}{G(1,\overline{\chi})\,y^s\,(1-e^{-ky})}$$
$$-ie^{\pi is/2}\left(\frac{k}{2\pi}\right)^{1-s}\int_0^\infty \frac{e^{-ky}G(-iky/2\pi,\overline{\chi})\,dy}{G(1,\overline{\chi})\,y^s\,(1-e^{-ky})}. \tag{12.5.12}$$

Replacing $m$ by $k-m$ in the definition of $G(z,\chi)$ shows that

$$e^{-ky}\,G\left(-\frac{iky}{2\pi},\overline{\chi}\right) = \chi(-1)\,G\left(\frac{iky}{2\pi},\overline{\chi}\right); \tag{12.5.13}$$

see Exercise 11. With the use of (12.5.8), (12.5.12) reduces to

$$L(s,\chi) = i\left(\frac{k}{2\pi}\right)^{1-s}\Gamma(1-s)\,L(1-s,\overline{\chi})\left\{\frac{e^{-\pi is/2}-\chi(-1)e^{\pi is/2}}{G(1,\overline{\chi})}\right\}. \tag{12.5.14}$$

Replace $s$ with $1-s$ and use (12.5.5) to convert (12.5.14) to (12.5.7): Exercise 11.

## 12.6 Other $L$-functions: algebraic and automorphic

Dirichlet's $L$-functions have been generalized or adapted in various directions. These various $L$-functions are central to some active areas of current research. In this section we describe a few examples, borrowed from Iwaniec and Sarnak [71].

As noted in [71], an $L$-function "is a type of generating function formed out of local data associated with either an arithmetic-geometric object ... or with an automorphic form." The term *local* here refers to the "primes" in some number field, e.g. the ordinary primes in the field $\mathbb{Q}$ of rationals. An automorphic form is a generalization of an automorphic function; see Section 7.8 and Chapter 17 for the latter.

We have already seen the first two examples; they are automorphic $L$-functions of degree 1. The third is an example of an automorphic $L$-function of degree 2. The fourth example comes from algebraic-geometric data.

1. The Riemann $\zeta$ function:

$$\zeta(s) = \prod_p (1-p^{-s})^{-1} = \sum_{n=1}^\infty n^{-s}.$$

2. Dirichlet $L$-functions:

$$L(s,\chi) = \prod_p (1-\chi(p)p^{-s})^{-1} = \sum_{n=1}^\infty \chi(p)p^{-s}.$$

3. An automorphic $L$-function of degree 2:

$$L(s,F) = \prod_{p}(1-\lambda(p)(Np)^{-s})^{-1}.$$

Here $p$ runs through the prime ideals of the field $\mathbb{Q}(\sqrt{-1})$ with $\lambda(\alpha) = (\alpha/|\alpha|)^4$, and $Np$ is the norm of $p$, while $F$ is a sum over the Gaussian integers:

$$F(z) = \sum_{m,n\in\mathbb{Z}} (m+in)^4 e^{2\pi i(m^2+n^2)z}, \quad z \in \mathbb{C}_+;$$

$F$ is an automorphic form for a certain subgroup of the modular group (the group of linear fractional transformations of the upper half plane that have integer coefficients and determinant 1), specifically the subgroup $\{f_A\}$ with the entry $A_{21}$ divisible by 4.

4. $L$-functions of elliptic curves over the rationals $\mathbb{Q}$:

$$L(s,E) = \prod_{p} L_p(s,E).$$

Here $E$ refers to a nonsingular curve

$$y^2 = x^3+ax+b, \qquad (12.6.1)$$

where $a$ and $b$ are integers. For all but finitely many primes $p$, the factor $L_p$ is

$$L_p(s,E) = \left(1-[p-N_E(p)]p^{-s-\frac{1}{2}}+p^{-2s}\right)^{-1},$$

where $N_E(p)$ is the number of integer solutions $(m,n) \mod p$ of

$$n^3 = m^3+am+b \mod p.$$

For the remaining primes the description of $L_p$ is more delicate.

There are many far-reaching conjectures concerning $L$-functions. One is the generalized Riemann hypothesis: all the non-trivial zeros of Dirichlet $L$-functions have real part $1/2$. Formulating other conjectures takes additional context. Quoting [71] again, referring to two conjectures about the relationship between the two types of $L$-functions, algebraic-geometric and automorphic, respectively: “it is expected that the latter set contains the former one, Shimura-Taniyama for special cases and Langlands in general.” The Shimura–Taniyama conjecture played a key role in Wiles’s proof of Fermat’s conjecture. The Langlands conjectures are a key part of the far-reaching “Langlands program.”

## Exercises

1. Suppose that $\chi : \mathbb{N} \to \{0\} \cup \{z : |z| = 1\}$ is multiplicative, has period $k$, and is not identically zero. Show that $\chi$ must have the properties (12.1.5) and (12.1.6).
2. Discuss $L(s,\chi)$ for the case $k = 2$.
3. Suppose that $k = 4$ and $\chi(1) = -\chi(-1) = 1$. Prove that $L(1,\chi) = \pi/4$.
4. Find all the characters for the groups $G_{10}$ and $G_{12}$ of remainders mod 10 and 12.
5. Suppose that $H$ is a subgroup of a finite group $G$. Define an equivalence relation in $G$ by $g_1 \sim g_2$ if $g_1^{-1} g_2$ belongs to $H$. Show that each equivalence class has the same number of elements, which shows that the order of $H$ divides the order of $G$ (Lagrange's theorem).
6. (a) Suppose that $g$ is an element of a finite group $G$. Show that there is a smallest integer $m > 0$ such that $g^m = 1$. This is called the *order* of the element $g$.

   (b) Use Exercise 5 to prove that $m$ divides the order of $G$.
7. The *Hurwitz zeta function* $\zeta(x,s)$ is defined by

$$\zeta(x,s) \;=\; \sum_{n=1}^{\infty} \frac{1}{(n+x)^s}, \qquad x > 0,\ \ \mathrm{Re}\, s > 1.$$

   Derive the integral representation

$$\zeta(x,s) \;=\; \frac{1}{\Gamma(s)} \int_{-\infty}^{\infty} \frac{e^{-xt}}{1-e^{-t}}\, t^{s-1}\, dt.$$

8. Suppose that $\chi$ is a Dirichlet character mod $k$.

   (a) Show that

$$L(s,\chi) \;=\; \frac{1}{m^s} \sum_{m=1}^{k-1} \chi(k)\, \zeta\left(\frac{k}{m}, s\right).$$

   (b) Relate this to (12.3.2) and (12.3.3).

   The remaining exercises are devoted to primitive Dirichlet characters and the functional equations for the associated $L$-functions.
9. Suppose that $\chi$ is a Dirichlet character mod $k$. Show that if $\chi$ is not primitive for $k$, then there is a unique smallest $\widetilde{k}$ with a character that induces $\chi$. This $\widetilde{m}$ is called the *conductor* of $\chi$.
10. The functional equation (12.5.7) is sometimes written in the form

$$\Lambda(1-s,\overline{\chi}) \;=\; \frac{i^a k^{1/2}}{G(1,\chi)} \Lambda(s,\chi), \tag{12.6.2}$$

   where

$$\Lambda(s,\chi) \;=\; \left(\frac{\pi}{k}\right)^{-(s+a)/2} \Gamma\left(\frac{s+a}{2}\right) L(s,\chi)$$

and $a = \frac{1}{2}[1 - \chi(-1)]$. Use the reflection formula (10.4.1) to derive (12.6.2) from (12.5.7).

11. Prove the unproved assertions in Section 12.5: (12.5.8), (12.5.9), (12.5.10), (12.5.11), (12.5.13), (12.5.7).

## Remarks and further reading

For much more information on this aspect of analytic number theory, see Apostol [8], [9], Dou and Zhang [37], Ireland and Rosen [70], and Moreno [100].

In connection with Section 12.6: in addition to Iwaniec and Sarnak [71], see Arthur [11], Gelbart [48], and Langlands [83].

# Chapter 13
# The Riemann hypothesis



In his famous paper on the zeta function [120], Riemann remarked that it is likely that all the non-trivial zeros of the zeta function lie on the line $\{s : \text{Re}\, s = \frac{1}{2}\}$. The "Riemann hypothesis" is the name that has been given to the assertion that this is the case, i.e. that all non-trivial zeros of $\zeta$ have real part $1/2$. Determining the truth of this assertion was one of the problems in Hilbert's famous list of outstanding mathematical problems (1900). The problem is still open at the time of this writing. It has (often) been called the greatest unsolved problem of mathematics. Among the reasons for this statement are the following:

- The hypothesis is equivalent to the statement about the degree of accuracy of Gauss's estimate for $\pi(x)$, the number of primes $\leq x$, as $x \to \infty$.
- The paper [120] in which Riemann made the statement contains some remarkable insights and assertions. Some of the assertions were proved only decades later, and some have not yet been fully verified (or disproven).
- Many leading analysts—Stieltjes, Hadamard, Hardy, Bohr, Landau, Polya, and Selberg, to name a few—have investigated Riemann's claims.
- There are close analogues of the hypothesis in other areas of mathematics. Some have been proved, others are open.
- Many feel that settling the hypothesis will require new methods that will have wider importance.

In this chapter we prove von Mangoldt's formula for a function $\psi(x)$ which is closely related to $\pi(x)$. This formula leads to a proof of the prime number theorem

$$\pi(x) \sim \frac{x}{\log x} \quad \text{as } x \to \infty,$$

and also to a proof of the equivalence of the Riemann hypothesis and the degree of accuracy of Gauss's empirical approximation to $\pi(x)$. Also included in this chapter is a brief discussion of Riemann's 1859 paper. A key tool is the inversion formula for the Mellin transform, which is proved in the last section.



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_13

## 13.1 Primes and zeros of the zeta function

The relation of the zeta function to the distribution of prime numbers is encoded in Euler's factorization (11.0.2), which we repeat here:

$$\zeta(s) = \prod_p (1-p^{-s})^{-1}, \quad \mathrm{Re}\, s > 1. \tag{13.1.1}$$

This factorization shows that $\zeta$ has no zeros in the half plane $\{s : \mathrm{Re}\, s > 1\}$. Therefore the principal branch of the logarithm of $1/\zeta$ is well defined there. Its derivative is

$$\begin{aligned} -\frac{\zeta'(s)}{\zeta(s)} &= \sum_p \frac{d}{ds}[\log(1-p^{-s})] = \sum_p \log p \, \frac{p^{-s}}{1-p^{-s}} \\ &= \sum_p \sum_{n=1}^{\infty} p^{-ns} \log p. \end{aligned} \tag{13.1.2}$$

The sum (13.1.2) can be written as a Stieltjes integral (see Section 1.8):

$$-\frac{\zeta'(s)}{\zeta(s)} = \int_0^\infty x^{-s}\, d\psi(x) = s\int_0^\infty x^{-s-1}\psi(x)\, dx, \tag{13.1.3}$$

for $\mathrm{Re}\, s > 1$, where

$$\psi(x) = \sum_{n=1}^{\infty} \sum_{p^n \le x} \log p. \tag{13.1.4}$$

(The second integral in (13.1.3) converges, since trivially $\psi(x) < x \log x$.)

Riemann's basic idea was to make use of two factorizations of the zeta function. One is Euler's factorization (13.1.1). The second factorization comes from the relation of $\zeta$ to the xi function:

$$\zeta(s) = \frac{2}{s(s-1)} \frac{1}{\Gamma(s/2)} \pi^{s/2} \xi(s). \tag{13.1.5}$$

In Section 11.5 it was shown that $\xi$ is an entire function that is symmetric about the line $\{z : \mathrm{Re}\, z = 1/2\}$. A consequence is that $\xi(\rho) = 0$ implies $\xi(1-\rho) = 0$. As shown in Section 8.5,

$$\sum_\rho \frac{1}{|\rho|^2} < \infty. \tag{13.1.6}$$

(This also follows from the result in Section 13.4.)

As shown in Section 8.5, $\xi$ has a factorization

$$\xi(s) = \xi(0) \prod_{\xi(\rho)=0} \left(1 - \frac{s}{\rho}\right), \tag{13.1.7}$$

which converges if the factors with $\rho$ and $1-\rho$ are paired, since

$$\frac{1}{s-\rho}+\frac{1}{s-1+\rho} = \frac{2s-1}{(s-\rho)(s-1+\rho)} = O(|\rho|^{-2}). \tag{13.1.8}$$

Thus the conditionally convergent product in (13.1.7) is taken to be

$$\lim_{N\to\infty}\prod_{|\rho|\le N}\left(1-\frac{s}{\rho}\right).$$

The factorization (10.1.4) of the gamma function gives

$$\frac{1}{\Gamma(s/2)} = \frac{s}{2}\prod_{n=1}^{\infty}\left(1+\frac{s}{2n}\right)\left(1+\frac{1}{n}\right)^{-s/2}. \tag{13.1.9}$$

Combining (13.1.5) with the factorizations (13.1.7) and (13.1.9), we find

$$-\frac{\zeta'(s)}{\zeta(s)} = \frac{1}{s-1}-\sum_{\rho}\frac{1}{s-\rho}-\sum_{n=1}^{\infty}\left[\frac{1}{s+2n}-\frac{1}{2}\log\left(1+\frac{1}{n}\right)\right]-\frac{1}{2}\log\pi. \tag{13.1.10}$$

The estimate (13.1.6) and the identity (13.1.8) show that the sum over $\rho$ in (13.1.10) converges if $\rho$ and $1-\rho$ are paired. The summands with respect to $n$ in (13.1.10) are

$$\begin{aligned}\frac{1}{s+2n}-\frac{1}{2}\log\left(1+\frac{1}{n}\right) &= \frac{1}{s+2n}-\frac{1}{2n}+r_n,\\ &= \frac{s}{2n(s+2n)}+r_n,\end{aligned}$$

where $r_n = O(n^{-2})$ is independent of $s$. (Here, and in many other places in this chapter, we use the estimate (1.5.2) for the principal branch of the logarithm.) Therefore the second sum in (13.1.10) also converges.

The resulting sum for $-\zeta'/\zeta$ extends from the half plane $\{s : \operatorname{Re} s > 1\}$ to the entire plane, yielding an entire meromorphic function with simple poles at $s=1$, $s=\rho$, and the even negative integers. In particular, $-\zeta'(0)/\zeta(0)$ is well defined. Subtracting its representation as a sum from that for $\zeta'/\zeta$, we obtain

$$\begin{aligned}-\frac{\zeta'(s)}{\zeta(s)} &= \frac{1}{s-1}+1-\sum_{\rho}\left[\frac{1}{s-\rho}+\frac{1}{\rho}\right]-\sum_{n=1}^{\infty}\left[\frac{1}{s+2n}-\frac{1}{2n}\right]-\frac{\zeta'(0)}{\zeta(0)}\\ &= \frac{s}{s-1}-\sum_{\rho}\frac{s}{\rho(s-\rho)}+\sum_{n=1}^{\infty}\frac{s}{2n(s+2n)}-\frac{\zeta'(0)}{\zeta(0)}.\end{aligned} \tag{13.1.11}$$

This series converges absolutely, apart from the poles at $s=1$, $s=\rho$, and $s=-2n$, $n=1,2,3,\ldots$.

## 13.2 von Mangoldt's formula for $\psi$

We now have two formulas for the logarithmic derivative $-\zeta'/\zeta$. One is (13.1.3), which involves the primes via the function

$$\psi(x) \;=\; \sum_{n=1}^{\infty}\sum_{p^n\le x}\log p.$$

The other is (13.1.11), which involves the non-trivial zeros of $\zeta$.

The formula (13.1.3) can be written as

$$-\frac{\zeta'(s)}{\zeta(s)} \;=\; s\,\widetilde{\psi}(s), \tag{13.2.1}$$

where

$$\widetilde{\psi}(s) \;=\; \int_0^{\infty} x^{-s-1}\psi(x)\,dx.$$

As noted above,

$$\psi(x) \;=\; \sum_{n=1}^{\infty}\sum_{p^n\le x}\log p \;<\; x\log x,$$

so $\widetilde{\psi}$ is well defined and holomorphic for $\operatorname{Re}s>1$. The function $\widetilde{\psi}$ would now be called the *Mellin transform* of $\psi$. The function $\psi$ can be recovered from $\widetilde{\psi}$ using the *Mellin inversion formula*

$$\psi(x) \;=\; \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty}\left[-\frac{\zeta'(s)}{s\,\zeta(s)}\right]x^s\,ds, \quad a>1. \tag{13.2.2}$$

In Section 13.7 we show that the integral (13.2.2) (taken in the sense specified in that section), recovers $\psi(x)$ at every point of continuity of $\psi$.

We see that the second formula for $\zeta'/\zeta$, (13.1.11), leads, via (13.2.2), to a formula for $\psi$:

$$\psi(x) \;=\; \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty}\left\{\frac{1}{s-1}-\sum_{\rho}\frac{1}{\rho\,(s-\rho)}+\sum_{n=1}^{\infty}\frac{1}{2n(s+2n)}-\frac{\zeta'(0)}{\zeta(0)\,s}\right\}x^s\,ds. \tag{13.2.3}$$

Assume for the moment that it is permissible to interchange integration and the summations. Each of the summands, except the last, has the form $-1/b(s-b)$, and the last is a constant time $1/s$. Thus, term-by-term, we are interested in

$$\frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty}\frac{x^s}{(s-b)}\,ds, \qquad \operatorname{Re}b\le 1<a. \tag{13.2.4}$$

If $0\le x<1$, the integrand is holomorphic in the half plane to the right of the line of integration. The integral around the rectangle bounded by the lines $\operatorname{Im}s=\pm h$, $\operatorname{Re}s=a$, and $\operatorname{Re}s=a+K$ is zero. The integral over the upper and lower sides is

dominated by

$$\frac{1}{h}\int_a^\infty x^s\,ds \;=\; \frac{x^a}{h\,|\log x|},$$

and the integral over the right side is dominated by (i.e. less than some constant times) $2hx^K$. Thus we may let $K \to \infty$ and conclude that

$$\left|\frac{1}{2\pi}\int_{a-ih}^{a+ih}\frac{x^s}{s-b}\,ds\right| \;=\; O(h^{-1}), \quad 0<x<1.$$

Suppose now that $x > 1$, and consider the integral around a rectangle bounded by the lines $\operatorname{Im} s = \pm h$, $\operatorname{Re} s = -K$ and $\operatorname{Re} s = a$. The integral over the left side is dominated by $2hx^{-K}$ for large $K$. The integrals over the upper and lower boundaries are each dominated, for large $h$, by

$$\frac{1}{h}\int_0^\infty e^{(a-t)\log x}\,dt \;=\; \frac{x^a}{h\log x},$$

independent of $K$. On the other hand, the value of the full integral around the rectangle is the residue at $s = b$, which is $x^b$. It follows that the integral (13.2.4) converges and

$$\frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty}\frac{x^s}{(s-b)}\,ds \;=\; \begin{cases} 0, & 0<x<1; \\ x^b, & x>1. \end{cases} \tag{13.2.5}$$

Formally, then, we have von Mangoldt's formula

$$\psi(x) \;=\; x - \sum_\rho \frac{x^\rho}{\rho} + \sum_{n=1}^\infty \frac{x^{-2n}}{2n} - \frac{\zeta'(0)}{\zeta(0)}, \qquad x>1. \tag{13.2.6}$$

The justification of the interchange of integration and summation that leads from (13.2.3) to (13.2.6) is taken up in Exercises 9–19.

## 13.3 The prime number theorem

A key step in the proof of the prime number theorem is to prove that $\psi(x) \sim x$ as $x \to \infty$, as suggested by the formula (13.2.6). The first step in the proof that $\psi(x) \sim x$ is to integrate (13.2.6):

$$\begin{aligned}\int_1^x \psi(t)\,dt &= \frac{x^2-1}{2} - \sum_\rho \frac{x^{\rho+1}}{\rho(\rho+1)} + O(1) - \frac{\zeta'(0)}{\zeta(0)}(x-1) \\ &= \frac{x^2}{2} - x^2\sum_\rho \frac{x^{\rho-1}}{\rho(\rho+1)} + O(x), \quad x>1.\end{aligned} \tag{13.3.1}$$

Since $\sum \rho^{-2}$ converges absolutely, and since $\operatorname{Re}(\rho-1) < 0$, the sum in the last line converges absolutely for each $x \geq 1$. Moreover each summand has limit 0 as $x \to \infty$.

This implies that

$$\int_0^x \psi(t)\,dt \sim \frac{x^2}{2} \text{ as } x \to \infty. \tag{13.3.2}$$

**Theorem 13.3.1.** *As $x \to \infty$, $\psi(x) \sim x$.*

*Proof:* According to (13.3.2), given $\varepsilon > 0$, there is an $N$ such that $x \geq N$ implies that

$$(1-\varepsilon)\frac{x^2}{2} \leq \int_0^x \psi(t)\,dt \leq (1+\varepsilon)\frac{x^2}{2}.$$

Then for $N \leq x < y$,

$$\int_x^y \psi(t)\,dt \leq (1+\varepsilon)\frac{y^2}{2} - (1-\varepsilon)\frac{x^2}{2} = \frac{y^2-x^2}{2} + \varepsilon\frac{x^2+y^2}{2},$$

and also

$$\int_x^y \psi(t)\,dt \geq (1-\varepsilon)\frac{y^2}{2} - (1+\varepsilon)\frac{x^2}{2} = \frac{y^2-x^2}{2} - \varepsilon\frac{x^2+y^2}{2}.$$

Since $\psi$ is non-decreasing,

$$\psi(x) \leq \frac{1}{y-x}\int_x^y \psi(t)\,dt \leq \psi(y),$$

so

$$(y-x)\,\psi(x) \leq \frac{y^2-x^2}{2} + \varepsilon\frac{x^2+y^2}{2};$$
$$(y-x)\,\psi(y) \geq \frac{y^2-x^2}{2} - \varepsilon\frac{x^2+y^2}{2}.$$

Take $y = \alpha x$, $\alpha > 1$. Then the preceding inequalities take the form

$$\frac{\psi(x)}{x} \leq \frac{(\alpha+1)}{2} + \varepsilon\frac{(\alpha^2+1)}{2(\alpha-1)} \leq \frac{(\alpha+1)}{2} + \frac{\varepsilon\alpha^2}{\alpha-1};$$
$$\frac{\psi(y)}{y} \geq \frac{(\alpha+1)}{2\alpha} - \varepsilon\frac{(\alpha^2+1)}{2\alpha(\alpha-1)} \geq \frac{(\alpha+1)}{2\alpha} - \frac{\varepsilon\alpha^2}{\alpha-1}.$$

Given $\delta > 0$, let $\alpha = 1+\delta$. For large $x$ we may take $\varepsilon\alpha^2/(\alpha-1) < \delta$, so that the preceding inequalities imply that

$$\frac{2+\delta}{2+2\delta} - \delta \leq \frac{\psi(x)}{x} \leq \frac{2+\delta}{2} + \delta.$$

Thus $\psi(x)/x \to 1$. □

The focus of Riemann's attention was the counting function

$$\pi(x) \;=\; \sum_{p\le x} 1 \;=\; \text{the number of primes } p \text{ less than or equal to } x. \tag{13.3.3}$$

Examining extensive tables of primes led Gauss to the conjecture that the density of primes near large $x$ was approximately $1/\log x$:

$$\pi(x) \;\sim\; \mathrm{Li}(x) \;=\; \mathrm{Li}(2) + \int_2^x \frac{dt}{\log t}. \tag{13.3.4}$$

(In general, $\mathrm{Li}(x)$, $x > 1$, is defined as a principal value integral

$$\mathrm{Li}(x) \;=\; \lim_{\delta\to 0+} \left\{ \int_0^{1-\delta} \frac{dt}{\log t} + \int_{1+\delta}^x \frac{dt}{\log t} \right\}; \tag{13.3.5}$$

see Exercise 1.)

**Lemma 13.3.2.** *The conjecture (13.3.4) is the same as the conjecture*

$$\pi(x) \;\sim\; \frac{x}{\log x}. \tag{13.3.6}$$

*Proof:* Integrating by parts, for $x > 2$

$$\int_2^x \frac{dt}{\log t} \;=\; \frac{x}{\log x} + \int_2^x \frac{dt}{(\log t)^2} - \frac{2}{\log 2}.$$

Now

$$\int_2^x \frac{dt}{(\log t)^2} = \int_2^{x^{1/2}} \frac{dt}{(\log t)^2} + \int_{x^{1/2}}^x \frac{dt}{(\log t)^2}$$
$$< \frac{x^{1/2}}{(\log 2)^2} + \frac{4x}{(\log x)^2} \;\sim\; \frac{x}{\log x}\cdot\frac{4}{\log x}$$

as $x \to \infty$. Therefore

$$\mathrm{Li}(x) \;=\; \frac{x}{\log x}\left\{1 + O\left(\frac{1}{\log x}\right)\right\}. \qquad \square$$

It is convenient at this point to introduce a new function. Formally, Gauss's approximation suggests that

$$\log x \, d\pi(x) \;=\; dx.$$

Thus the integrated version of (13.3.4) would be

$$\vartheta(x) \;\equiv\; \sum_{p\le x} \log p \;=\; \int_0^x \log t \, d\pi(t) \;\sim\; x. \tag{13.3.7}$$

The next step in proving (13.3.4) is to prove (13.3.7).

**Lemma 13.3.3.** *As $x \to \infty$, $\vartheta(x) \sim x$.*

*Proof:* Let us relate $\vartheta$ to $\psi$, whose asymptotic behavior we know. Clearly

$$\vartheta(x) \;\le\; \psi(x) \;=\; \sum_{n=1}^{\infty} \sum_{p^n \le x} \log p.$$

On the other hand

$$\psi(x) \;=\; \vartheta(x) + \vartheta(x^{1/2}) + \vartheta(x^{1/3}) + \cdots + \vartheta(x^{1/n}),$$

where $x^{1/(n+1)} < 2$. Thus the number of summands is less than $\log x / \log 2$, so

$$\vartheta(x) \;\le\; \psi(x) \;\le \vartheta(x) + \frac{\log x}{\log 2}\, \vartheta(x^{1/2}). \tag{13.3.8}$$

Since $\psi(x) \sim x$,

$$\vartheta(x) \;\le\; \psi(x) \;\le\; \vartheta(x) + O(x^{1/2} \log x), \tag{13.3.9}$$

which implies $\vartheta(x) \sim x$. □

Now

$$\vartheta(x) \;=\; \int_0^x \log t \, d\pi(t) \;=\; \pi(x) \log x - \int_0^x \frac{\pi(t)}{t}\, dt.$$

But $\pi(t)/t$ is bounded and $\pi$ is non-decreasing, so

$$\int_0^x \frac{\pi(t)}{t}\, dt \;=\; \int_0^s \frac{\pi(t)}{t}\, dt + \int_s^x \frac{\pi(t)}{t}\, dt \;\le\; O(s) + \pi(x) \log\left(\frac{x}{s}\right).$$

Taking $s = x/\log x$ gives

$$\vartheta(x) \;=\; \pi(x) \log x + O(x/\log x) + \pi(x) O(\log\log x).$$

Combining this with Lemma 13.3.3, we obtain the prime number theorem.

**Theorem 13.3.4.** (Prime Number Theorem) *As $x \to \infty$,*

$$\pi(x) \;\sim\; \frac{x}{\log x}.$$

## 13.4 Density of the zeros

A key ingredient in the detailed analysis of the relationship between the zeros $\rho$ and the accuracy of Gauss's estimate for $\pi(s)$ is more information about the distribution of the $\rho$. Riemann stated a very precise estimate of the density of the zeros $\rho$ of the zeta function. A weaker result was proved by von Mangoldt [96] nearly a half

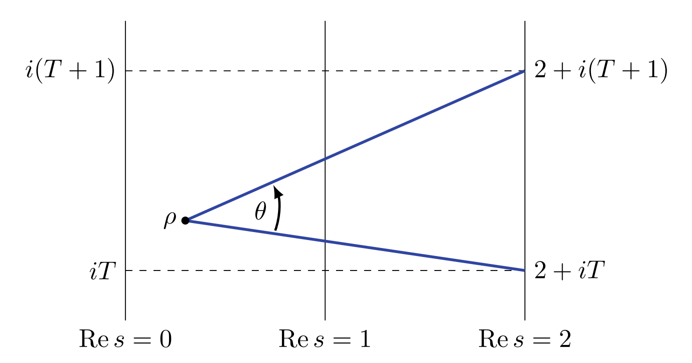



**Fig. 13.1** Change in argument of $s-\rho$

century later. Note that because of the pairing $\rho \leftrightarrow 1-\rho$, it is enough to consider the density for $\mathrm{Im}\,\rho > 0$.

**Theorem 13.4.1.** (von Mangoldt) *The asymptotic density of the zeros $\rho$ of $\xi$ is at most $2\log x$, in the sense that for $T \geq T_0$, the number of roots $\rho$ such that $T \leq \mathrm{Im}\,\rho \leq T+1$ is at most $2\log T$.*

*Proof:* The proof involves integrating $\xi'/\xi$ from $2+iT$ to $2+i(T+1)$. According to (13.1.7), the integrand is the series

$$\sum_{\rho} \frac{d}{ds} \log\left(1-\frac{s}{\rho}\right) = \sum_{\rho} \frac{1}{s-\rho},$$

which is conditionally convergent: see (13.1.8). Integrating a single term gives the change in $\log(s-\rho)$ from $s = 2+iT$ to $s = 2+i(T+1)$. The imaginary part is the change in angle, and is always positive. If $T \leq \mathrm{Im}\,\rho \leq T+1$, then, since $\mathrm{Re}\,\rho > 0$, this change of angle is at least $\tan^{-1}(1/2)$; see Figure 13.1.

Therefore if there are $n(T)$ zeros with imaginary part between $T$ and $T+1$,

$$n(T)\tan^{-1}(1/2) \leq \mathrm{Im}\left\{\int_{2+iT}^{2+i(T+1)} \frac{\xi'(s)}{\xi(s)}\, ds\right\} \tag{13.4.1}$$

$$= \mathrm{Im}\left\{\log \frac{\xi(2+i[T+1])}{\xi(2+iT)}\right\}. \tag{13.4.2}$$

According to the definition (13.1.5),

$$\log \xi(s) = \log \Gamma\left(\frac{1}{2}s\right) + \log s + \log(1-s)$$
$$- \log 2 - \frac{s}{2}\log \pi + \log \zeta(s), \tag{13.4.3}$$

so letting $S = 2+iT$ we have $|S| \to \infty$ as $T \to \infty$, and

$$\begin{aligned}
\log \frac{\xi(2+i[T+1])}{\xi(2+iT)} &= \log \xi(S+i) - \log \xi(S) \\
&= \log \Gamma\left(\frac{1}{2}(S+i)\right) - \log \Gamma\left(\frac{1}{2}S\right) + \log \frac{S+i}{S} + \log \frac{1-S-i}{1-S} \\
&\quad - \frac{i}{2}\log \pi + \log \frac{\zeta(S+i)}{\zeta(S)} \\
&= \log \Gamma\left(\frac{1}{2}(S+i)\right) - \log \Gamma\left(\frac{1}{2}S\right) + \log\left(1+\frac{i}{S}\right) + \log\left(\frac{1-S-i}{1-S}\right) \\
&\quad - \frac{i}{2}\log \pi + \log \frac{\zeta(S+i)}{\zeta(S)} \\
&= \log \Gamma\left(\frac{1}{2}(S+i)\right) - \log \Gamma\left(\frac{1}{2}S\right) + \log \frac{\zeta(S+i)}{\zeta(S)} + O(1).
\end{aligned}$$

Now $|1/p^S| = |1/p^{S+i}| = 1/p^2$, so the product formula (13.1.1) shows that

$$\log \frac{\zeta(S+i)}{\zeta(S)} = \sum_p \left[\log\left(1-\frac{1}{p^S}\right) - \log\left(1-\frac{1}{p^{S+i}}\right)\right] = O\left(\sum_p \frac{1}{p^2}\right) = O(1).$$

Finally, the estimate (10.5.8) gives

$$\begin{aligned}
&\log \Gamma\left(\frac{(S+i)}{2}\right) - \log \Gamma\left(\frac{S}{2}\right) \\
&\sim \frac{1}{2}\left[(S+i-1)\log \frac{S+i}{2} - (S+i) - (S-1)\log \frac{S}{2} + S\right] \\
&\sim \frac{S-1}{2}\log\left(1+\frac{i}{S}\right) + \frac{i}{2}\log\left(\frac{1}{2}[S+i]\right) \\
&\sim \frac{i}{2}\log(S+i) \sim \frac{i}{2}\log T.
\end{aligned}$$

It follows from this and from (13.4.2) that the number of zeros with imaginary part between $T$ and $T+1$ is less than

$$\frac{1}{2\tan^{-1}(1/2)}\log T + O(1) \le 2\log T$$

as $T \to \infty$. □

The density estimate leads to another proof of convergence of $\sum(1/|\rho|^2)$. The number of roots $\rho$ with $|\rho| \sim n^2$ is dominated, for large $n$, by a multiple of $\log n$ , so the sum $\sum|\rho|^{-2}$ is dominated by a multiple of $\sum \log n/n^2$.

## 13.5 The Riemann hypothesis and Gauss's approximation

How accurate can the approximation (13.3.4) be? Taken together, the two theorems in this section show that the Riemann hypothesis is *equivalent* to a certain degree of accuracy of Gauss's estimate (13.3.4).

**Theorem 13.5.1.** *Suppose that for each* $\varepsilon > 0$*,*

$$\pi(x) \;=\; \mathrm{Li}(x) + O(x^{\frac{1}{2}+\varepsilon}) \quad \text{as} \;\; x \to \infty. \tag{13.5.1}$$

*Then the Riemann hypothesis is true: all non-trivial zeros of the zeta function lie on the line* $\{s : \mathrm{Re}\, s = \frac{1}{2}\}$*.*

We begin with a lemma relating this estimate to an estimate for $\psi$.

**Lemma 13.5.2.** *Suppose that (13.5.1) is true for each* $\varepsilon > 0$*. Then for each* $\varepsilon > 0$*,*

$$\psi(x) - x \;=\; O(x^{\frac{1}{2}+\varepsilon}) \quad \text{as} \;\; x \to \infty. \tag{13.5.2}$$

*Proof:* In view of the estimate (13.3.9) it is enough to prove that

$$\vartheta(x) - x \;=\; O(x^{\frac{1}{2}+\varepsilon}) \quad \text{as} \;\; x \to \infty. \tag{13.5.3}$$

Now $dx = \log x\, d\,\mathrm{Li}(x)$, so given $a > 1$,

$$\begin{aligned} \vartheta(x) - x &= \int_a^x \log t\, d[\pi(t) - \mathrm{Li}(t)] - [\vartheta(a) - a] \\ &= \log x\,[\pi(x) - \mathrm{Li}(x)] - \int_a^x \frac{\pi(t) - \mathrm{Li}(t)}{t}\, dt - [\vartheta(a) - a]. \end{aligned}$$

Given $\varepsilon > 0$, the constant $a$ can be chosen so that $x \ge a$ implies

$$|\pi(x) - \mathrm{Li}(x)| \le x^{\frac{1}{2}+2\varepsilon},$$

so

$$\begin{aligned} |\vartheta(x) - x| &\le x^{\frac{1}{2}+2\varepsilon} \log x + \int_a^x t^{-\frac{1}{2}+2\varepsilon}\, dt + \text{constant} \\ &\le x^{\frac{1}{2}+3\varepsilon} \end{aligned}$$

for large $x$. $\square$

*Proof of Theorem* 13.5.1. The function $-(1-s)\zeta(s)$ is entire. By (13.1.3), its logarithmic derivative is

$$\begin{aligned}\frac{1}{s-1}+\frac{\zeta'(s)}{\zeta(s)} &= \frac{1}{s-1}-s\int_1^\infty \psi(x)x^{-s-1}\,dx\\ &= \int_1^\infty x^{-s}\,dx - s\int_1^\infty \psi(x)x^{-s-1}\,dx\\ &= -1+s\int_1^\infty [x-\psi(x)]\,x^{-s-1}\,dx. \qquad (13.5.4)\end{aligned}$$

Assume (13.5.1). By Lemma 13.5.2,

$$x-\psi(x) \;=\; O(x^{\frac{1}{2}+\varepsilon}),$$

so the integral in the last line of (13.5.4) converges in the half plane $\{s : \mathrm{Re}\, s > \frac{1}{2}+\varepsilon\}$. This shows that the derivative of the logarithm of $(s-1)\zeta(s)$ is holomorphic in this half plane, so $\zeta(s)$ has no zeros in this half plane. We know that the set of non-trivial zeros of $\zeta$ is symmetric about the line $\mathrm{Re}\, s = 1/2$, so there are also no zeros in the strip $\{0 < \mathrm{Re}\, s < 1/2-\varepsilon\}$. By assumption this is true for every $\varepsilon > 0$. Thus the assumption (13.5.1) implies the Riemann hypothesis. □

The converse of this theorem, proved by von Koch [78], is deeper. It relies on the full strength of results of the preceding sections.

**Theorem 13.5.3.** (von Koch) *If the non-trivial zeros of the zeta function all lie on the line* $\{s : \mathrm{Re}\, s = \frac{1}{2}\}$, *then*

$$\pi(x)-\mathrm{Li}(x) \;=\; O(x^{1/2}\log x). \qquad (13.5.5)$$

*Proof:* We start with the estimate

$$\begin{aligned}\psi(x) \le \int_x^{x+1}\psi(t)\,dt &= \int_0^{x+1}\psi(t)\,dt - \int_0^x \psi(t)\,dt\\ &= \frac{(x+1)^2-x^2}{2} - \sum_\rho \frac{(x+1)^{\rho+1}-x^{\rho+1}}{\rho(\rho+1)} + O(1)\\ &\le x + \sum_\rho \left|\frac{(x+1)^{\rho+1}-x^{\rho+1}}{\rho(\rho+1)}\right| + O(1). \qquad (13.5.6)\end{aligned}$$

Similarly

$$\begin{aligned}\psi(x) &\ge \int_{x-1}^x \psi(t)\,dt\\ &\ge x - \sum_\rho \left|\frac{x^{\rho+1}-(x-1)^{\rho+1}}{\rho(\rho+1)}\right| + O(1). \qquad (13.5.7)\end{aligned}$$

Suppose that each $\rho$ has real part $\frac{1}{2}$. Then $|x^\rho| = x^{1/2}$, so for $x \ge 1$,

$$\left|\frac{(x+1)^{\rho+1}-x^{\rho+1}}{\rho(\rho+1)}\right| \le \frac{2(x+1)^{3/2}}{\rho(\rho+1)} \;\le\; \frac{2(2x)^{3/2}}{\gamma^2}, \quad \gamma = \mathrm{Im}\,\rho. \qquad (13.5.8)$$

We also want the estimate, for $x \geq 1$,

$$\left| \frac{(x+1)^{\rho+1} - x^{\rho+1}}{\rho(\rho+1)} \right| = \left| \int_x^{x+1} \frac{t^\rho}{\rho}\, dt \right| \leq \frac{(2x)^{1/2}}{|\gamma|}, \quad \gamma = \operatorname{Im}\rho, \tag{13.5.9}$$

which is stronger than (13.5.8) for $x > |\gamma|$.

Suppose now that $x$ is larger than the constant $T_0$ of Theorem 13.4.1. Combining the estimates (13.5.6), (13.5.8), and (13.5.9), and continuing to denote $\operatorname{Im}\rho$ by $\gamma$, we have

$$\psi(x) \;\leq\; x + (2x)^{1/2} \sum_{|\gamma|<T_0} \frac{1}{|\gamma|} + (2x)^{1/2} \sum_{T_0 \leq |\gamma| < x} \frac{1}{|\gamma|} + 2(2x)^{3/2} \sum_{|\gamma| \geq x} \frac{1}{\gamma^2}.$$

The first sum is a constant multiple of $x^{1/2}$. Theorem 13.4.1 shows that the second and third sums are dominated by

$$\int_{T_0}^x \frac{\log t}{t}\, dt + \int_x^\infty \frac{\log t}{t^2}\, dt \;=\; \frac{1}{2}\left[(\log x)^2 - (\log T_0)^2\right] + \frac{\log x}{x} + \frac{1}{x}.$$

It follows from these estimates and (13.5.6), (13.5.7) that

$$\psi(x) \;=\; x + O\left(x^{1/2}(\log x)^2\right).$$

Then (13.3.9) gives the same estimate for $\vartheta$:

$$\vartheta(x) \;=\; x + O\left(x^{1/2}(\log x)^2\right). \tag{13.5.10}$$

Now $d\vartheta \;=\; \log x\, d\pi$ so (13.5.10) implies that

$$\begin{aligned}\pi(x) - \operatorname{Li}(x) + \operatorname{Li}(2) &= \int_2^x \frac{1}{\log t}\, d(\vartheta(t) - t)\\ &= \frac{1}{\log x}[\vartheta(x) - x] + \int_2^x [\vartheta(t) - t]\frac{dt}{t(\log t)^2}.\end{aligned}$$

In view of (13.5.10), the first summand on the right is $O(x^{1/2}\log x)$ and the integrand in the second summand is $O(t^{-1/2})$, so

$$\pi(x) - \operatorname{Li}(x) \;=\; O(x^{1/2}\log x). \qquad \square$$

## 13.6 Riemann's 1859 paper

In this section we outline Riemann's work on this subject, but with updated and modified notation. Riemann works with $\Pi(s) \;=\; \Gamma(s+1)$ and denotes the counting function $\pi(x)$ by $F(x)$, for example.

Riemann begins by noting Euler's factorization (13.1.1). He derives the equation (11.2.4) in order to extend $\zeta$, noting that the extension has a single, simple, pole at $s=1$. He uses this formula to derive the functional equation (11.2.5). He introduces the function

$$\sigma(s) \;=\; \Gamma\left(\frac{s}{2}\right)\pi^{-s/2}\,\zeta(s) \tag{13.6.1}$$

and notes that this function is unchanged under $s \to 1-s$. He then proceeds to a second representation of $\sigma$ in terms of a theta function. Specifically he sets

$$\theta(x) \;=\; \sum_{n=1}^{\infty} e^{-n^2\pi x}$$

and shows that

$$\sigma(s) \;=\; \int_0^\infty \theta(x)\,x^{s/2-1}\,dx; \tag{13.6.2}$$

see Exercise 4. He uses Jacobi's identity

$$\sum_{-\infty}^{\infty} e^{-n^2\pi x} \;=\; 1+2\theta(x) \;=\; x^{-1/2}\left[1+2\theta\left(\frac{1}{x}\right)\right] \tag{13.6.3}$$

to derive the equation

$$\begin{aligned}
\sigma(s) &= \int_1^\infty \theta(x)\,x^{s/2-1}\,dx + \int_0^1 \theta\left(\frac{1}{x}\right)x^{(s-3)/2}\,dx \\
&\quad + \frac{1}{2}\int_0^1 \left(x^{(s-3)/2} - x^{s/2-1}\right)dx \\
&= \frac{1}{s(s-1)} + \int_1^\infty \theta(x)\left(x^{s/2-1} + x^{-(1+s)/2}\right)dx.
\end{aligned}$$

(Jacobi's identity (13.6.3) follows from the Poisson summation formula. See Exercises 17–19 of Chapter 18.)

Riemann then introduces the function $\xi$, so that the previous equation becomes

$$\begin{aligned}
\xi\left(\frac{1}{2}+it\right) &= \frac{1}{2} - \left(t^2+\frac{1}{4}\right)\int_1^\infty \theta(x)x^{-3/4}\cos\left(\frac{t}{2}\log x\right)dx \\
&= 4\int_1^\infty \frac{d}{dx}\left\{x^{3/2}\,\theta'(x)\right\}x^{-1/4}\cos\left(\frac{t}{2}\log x\right)dx,
\end{aligned} \tag{13.6.4}$$

and notes that $\xi(\frac{1}{2}+it)$ is entire, and all zeros have $|\mathrm{Im}\,t| \le \frac{1}{2}$. He also states that the number of zeros $\rho$ with $\mathrm{Im}\,\rho$ between 0 and $T$ is approximately

$$\frac{T}{2\pi}\log\frac{T}{2\pi} - \frac{T}{2\pi},$$

and it is very likely ("es ist sehr warscheinlich") that all the zeros $\rho$ have real part $1/2$.

At this point Riemann makes a statement that is equivalent to the factorization (13.1.7)

$$\xi(s) = \xi(0)\prod_{\rho}\left(1-\frac{s}{\rho}\right),$$

an assertion that was first proved by Hadamard more than 30 years later, see Section 8.5.

Riemann now introduces the counting function, denoted here by $\pi$, and a related function

$$J(x) = \sum_{n=1}^{\infty}\sum_{p^n\le x}\frac{1}{n} = \pi(x)+\tfrac{1}{2}\pi(x^{1/2})+\tfrac{1}{3}\pi(x^{1/3})+\tfrac{1}{4}\pi(x^{1/4})+\dots,$$

and derives the identity

$$\frac{\log\zeta(s)}{s} = \int_0^{\infty} J(x)x^{-s-1}\,dx,$$

which is equivalent to (13.1.3). Riemann identifies this Mellin transform as being equivalent to a Fourier transform and inverts it, half a century before Mellin's paper. The result is

$$J(x) = \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty}\frac{\log\zeta(s)}{s}x^s\,ds.$$

Riemann inserts into this formula the representation of $\log\zeta$ that corresponds to the formula (13.1.10). After a page of manipulations, he arrives at a formula

$$J(x) = \mathrm{Li}(x) - \sum_{\mathrm{Im}\,\rho>0}\left[\mathrm{Li}(x^{\rho})+\mathrm{Li}(x^{1-\rho})\right] + \int_x^{\infty}\frac{1}{x^2-1}\cdot\frac{dx}{x\log x}+\log\xi\left(\frac{1}{2}\right), \tag{13.6.5}$$

where the $\rho, 1-\rho$ are the non-trivial zeros of $\zeta$. This formula is the analogue of (13.2.6). It was proved by von Mangoldt [96] some forty years after Riemann's work.

## 13.7 Inverting the Mellin transform of $\psi$

Recall that the non-negative function $\psi$ vanishes for $0\le x<2$ and satisfies the trivial estimate $\psi(x)\le x\log x$. Therefore the Mellin transform

$$\widetilde{\psi}(s) = \int_0^{\infty}\psi(x)x^{-s-1}\,ds$$

is holomorphic for $\mathrm{Re}\,s>1$, and is bounded on each line $\mathrm{Re}\,s=a$, $a>1$. We claim that at each point of continuity $x$ of $\psi$,

$$\begin{aligned}\psi(x) &= \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} \widetilde{\psi}(s)\,x^s\,ds \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{\psi}(a+it)\,x^{a+it}\,dt. \end{aligned} \tag{13.7.1}$$

The meaning of the integral is not clear, based simply on the boundedness of $\widetilde{\psi}$. We define it by introducing a convergence factor, and taking the integral to be the limit:

$$\lim_{\varepsilon\to 0+}\left\{\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\varepsilon|t|}\,\widetilde{\psi}(a+it)\,x^{a+it}\,dt\right\}. \tag{13.7.2}$$

Our objective is to show that this limit exists and equals $\psi(x)$ at each point of continuity of $\psi$. The convergence factor allows a change in the order of integration:

$$\begin{aligned}&\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\varepsilon|t|}\,\widetilde{\psi}(a+it)\,x^{a+it}\,dt \\ &\quad= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\varepsilon|t|}x^{a+it}\left\{\int_0^{\infty}\psi(y)\,y^{-a-it-1}\,dy\right\}dt \\ &\quad= \frac{1}{2\pi}\int_0^{\infty}\psi(y)\left(\frac{x}{y}\right)^a\left\{\int_{-\infty}^{\infty} e^{-\varepsilon|t|}\left(\frac{x}{y}\right)^{it}dt\right\}\frac{dy}{y}. \end{aligned} \tag{13.7.3}$$

Let $x/y = e^{-\varepsilon u}$. Then the inner integral in the last line is

$$\begin{aligned}\int_{-\infty}^{\infty} e^{-\varepsilon|t|-i\varepsilon tu}\,dt &= \int_0^{\infty}\left[e^{-\varepsilon t(1-iu)} + e^{-\varepsilon t(1+iu)}\right]dt \\ &= \frac{1}{\varepsilon(1-iu)} + \frac{1}{\varepsilon(1+iu)} = \frac{2}{\varepsilon(1+u^2)}. \end{aligned}$$

Since $y = xe^{\varepsilon u}$, it follows that (13.7.3) is

$$\frac{1}{\pi}\int_0^{\infty}\psi(xe^{\varepsilon u})\,e^{-\varepsilon a u}\,\frac{du}{1+u^2}. \tag{13.7.4}$$

Now

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{du}{1+u^2} = 1,$$

(Exercise 6), so it follows from (13.7.4) that the limit (13.7.2) is $\psi(x)$ wherever $\psi$ is continuous, Exercise 7.

## Exercises

1. Prove that the limit (13.3.5) exists.

2. Suppose that for each $\varepsilon > 0$, $|\pi(x) - \mathrm{Li}(x)| \le x^{\alpha+\varepsilon}$ for any sufficiently large $x$, where $\frac{1}{2} < \alpha < 1$. Prove that all non-trivial zeros of $\zeta$ are in the strip $1 - \alpha \le \mathrm{Re}\, s \le \alpha$.
3. Suppose that all the non-trivial zeros of $\zeta$ lie in the strip $1 - \alpha \le \mathrm{Re}\, s \le \alpha$, where $\frac{1}{2} < \alpha < 1$. Prove that $|\pi(x) - \mathrm{Li}(x)| = O(x^{\alpha+\varepsilon})$.
4. Use the integral form of the gamma function and write $\zeta$ as a sum to derive (13.6.2).
5. Use Jacobi's identity (13.6.3) to derive (13.6.4).
6. Use the residue calculus to prove that $\int_{-\infty}^{\infty}(1+s^2)^{-1} = \pi$.
7. Assuming that $\psi$ is continuous at $x$, fill in the details in the assertion that the limit (13.7.1) is $\psi(x)$.
8. Suppose that $\psi$ has a jump discontinuity at $x$. Prove that

$$\frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} \widetilde{\psi}(s)\, x^s\, ds \;=\; \frac{1}{2}\left[\lim_{\varepsilon\to 0+} \psi(x+\varepsilon) - \lim_{\varepsilon\to 0+} \psi(x-\varepsilon)\right].$$

9. This is the first in a series of exercises that justify the interchange of summation and integration in the proof of von Mangoldt's formula (13.2.6).

   Recall that if $\rho$ is a zero of $\xi$ then so is $1-\rho$, so that in estimating sums it is enough to sum over $\mathrm{Im}\,\rho \ge 0$.

   Show that the two series in (13.1.11) converge uniformly on bounded intervals of the line $\{s : \mathrm{Re}\, s = a > 1\}$, so that the resulting expression for

$$-\frac{1}{2\pi i}\int_{a-ih}^{a+ih} \frac{\zeta'(s)}{s\,\zeta(s)}\, x^s\, ds$$

   can be integrated term-by-term.
10. Show that for each $x > 0$, $a > 0$, $h > 0$,

$$\left|\int_{a-ih}^{a+ih} \frac{x^s}{s}\, ds\right| \;\le\; \frac{4x^a}{a\log x}.$$

   (Hint: $x^s = (1/\log x)\,d(x^s)/ds$, integrate by parts.)
11. Use Exercise 10 to prove

$$\left|\int_{a-ih}^{a+ih} \frac{x^s}{s+2n}\, ds\right| \;\le\; \frac{2x^a}{n\log x}.$$

12. Use Exercises 9 and 11 to show that

$$\frac{1}{2\pi i}\lim_{h\to\infty}\int_{a-ih}^{a+ih} \sum_{n=1}^{\infty} \frac{x^s}{2n(2n+s)}\, ds \;=\; \sum_{n=1}^{\infty} \frac{x^{-2n}}{2n}$$

   for $x > 1$, and the limit is 0 for $0 < x < 1$.
13. Suppose $a > 0$ and $0 < b < h$. Show that

$$\left|\int_{a+ib}^{a+ih} \frac{x^s}{s}\,ds\right| \leq K\frac{x^a}{\log x}\cdot\frac{1}{a+b}.$$

($K$ can be taken to be $4\sqrt{2}$.)

14. Suppose that $a > 1$ and that $\rho = \sigma + i\tau$ is a zero of $\xi$ with $\tau > h > 0$. Use the identity
$$\int_{a-ih}^{a+ih} \frac{x^s}{s-\rho}\,ds \;=\; x^\rho \int_{a-\rho-ih}^{a-\rho+ih} \frac{x^t}{t}\,dt$$
and Exercise 13 to prove that
$$\left|\int_{a-ih}^{a+ih} \frac{x^s}{s-\rho}\right| \;\leq\; K\frac{x^a}{\log x}\cdot\frac{1}{c+\tau-h},$$
where $c = a - \sigma > a - 1 > 0$.

15. Assume that $h > T_0$, the constant in Theorem 13.4.1. Deduce from Exercise 14 that the sum
$$\sum_{|\mathrm{Im}\,\rho|>h}\left|\int_{a-ih}^{a+ih} \frac{x^s}{\rho(s-\rho)}\,ds\right|$$
is dominated by
$$\begin{aligned}\frac{x^a}{\log x}\int_h^\infty \frac{\log\tau}{\tau(\tau-h+c)}\,d\tau &= \frac{x^a}{\log x}\int_0^\infty \frac{\log(t+h)}{(t+h)(t+c)}\,dt\\ &\leq \frac{x^a}{\log x}\int_0^\infty \frac{1}{(t+h)^{1/2}(t+c)}\,dt\\ &\leq \frac{x^a}{\log x}\int_0^\infty \frac{1}{h^{1/4}t^{1/4}(t+c)}\,dt\end{aligned}$$
and conclude that this sum converges to zero as $h\to\infty$.

16. Use the argument that leads to (13.2.5) to show that
$$\left|\frac{1}{2\pi i}\int_{a-ih}^{a+ih} \frac{x^s}{s}\,ds - 1\right| \;\leq\; \frac{x^a}{\pi h\log x}.$$

17. Suppose that $\rho = \sigma + i\tau$, $\tau > 0$, is a zero of $\xi$.

(a) Show that
$$\frac{1}{2\pi i}\int_{a-ih}^{a+ih} \frac{x^s}{\rho(s-\rho)}\,ds - \frac{x^\rho}{\rho} \;=\; \frac{x^\rho}{\rho}\left\{\frac{1}{2\pi i}\int_{a-\rho-ih}^{a-\rho+ih} \frac{x^t}{t}\,dt - 1\right\}$$
and, setting $c = a - \sigma > 1/2$,
$$\frac{1}{2\pi i}\int_{a-\rho-ih}^{a-\rho+ih} \frac{x^t}{t}\,dt - 1 \;=\; \left\{\frac{1}{2\pi i}\int_{c-i(h+\tau)}^{c+i(h+\tau)} \frac{x^t}{t}\,dt - 1\right\} - \frac{1}{2\pi i}\int_{c+i(h-\tau)}^{c+i(\tau+h)} \frac{x^t}{t}\,dt.$$

(b) Use Exercise 15 to show that the first expression on the right in the last line of part (a) is dominated by

$$\frac{x^{a-\sigma}}{h\log x}.$$

(c) Use Exercise 13 to show that the second expression on the right in the last line of part (a) is dominated by

$$\frac{x^{a-\sigma}}{(c+h-\tau)\log x}.$$

18. Use Exercises 16 and 17 to show that, for $c = a-1 > 0$,

$$\sum_{0<|\mathrm{Im}\,\rho|\le h}\left\{\frac{1}{2\pi i}\int_{a-ih}^{a+ih}\frac{x^\rho}{\rho(s-\rho)}\,ds - \frac{x^\rho}{\rho}\right\}$$
$$= \sum_{0<|\mathrm{Im}\,\rho|\le h}\frac{x^\rho}{\rho}\frac{1}{2\pi i}\int_{a-ih}^{a+ih}\left\{\frac{x^\rho}{\rho(s-\rho)}\,ds - 1\right\}$$

is dominated by

$$\frac{x^a}{\log x}\sum_{0\le \mathrm{Im}\,\rho\le h}\left\{\frac{1}{\mathrm{Im}\,\rho\,(h+\mathrm{Im}\,\rho)} + \frac{1}{\mathrm{Im}\,\rho\,(c+h)}\right\}. \tag{13.7.5}$$

19. In view of Exercises 9, 12, and 15, we can complete the proof of von Mangoldt's formula (13.2.6), by showing that (13.7.5) has limit zero as $h \to \infty$. Individual summands have limit zero, so we may ignore the terms with $\mathrm{Im}\,\rho < T_0$, the constant in Theorem 13.4.1, and estimate the remaining sums by

$$\int_{T_0}^{h}\frac{\log\tau}{\tau(\tau+h)}\,d\tau + \int_{T_0}^{h}\frac{\log\tau}{\tau(c+h-\tau)}\,d\tau.$$

Use the identities

$$\frac{1}{\tau(h+\tau)} = \frac{1}{h}\left(\frac{1}{\tau} - \frac{1}{h+\tau}\right); \qquad \frac{1}{\tau(c+h-\tau)} = \frac{1}{c+h}\left(\frac{1}{\tau} - \frac{1}{c+h-\tau}\right)$$

to show that both integrals are dominated by $(\log h)^2/h$.

## Remarks and further reading

The book by Edwards [42] contains an extensive discussion, with proofs, of Riemann's paper, von Mangoldt's proof of Riemann's formula, and subsequent developments, including the proofs by Hadamard [56] and de la Vallée Poussin [35] of the prime number theorem, as well as methods of computing and locating zeros of $\zeta$. The book edited by Borwein et al. [27] contains a wealth of related material, including essays by experts and reprintings of many of the important papers in the subject.

There have been many proofs of the prime number theorem since those of Hadamard and de la Vallée Poussin, whose papers are among those that are included in [27]. Proofs that are "elementary," in the sense of not using complex variables, were found by Selberg and by Erdős. The Erdős and Selberg papers, as well as a short proof of Newman [107] that does use complex analysis, are also included in [27]. For Newman's proof, see also the exposition by Zagier [147].

# Chapter 14
# Elliptic functions and theta functions




The trigonometric functions are the basic functions that are periodic with respect to a translation of the plane $\mathbb{C}$. An important class of complex functions is *doubly periodic*: periodic with respect to two sets of translations. This chapter presents the general theory of such functions, and Jacobi's construction via theta functions. The following two chapters, which are independent of each other, present constructions due to Jacobi and to Weierstrass, respectively.

Elliptic functions play, and have played, a key role in many developments in number theory and algebraic geometry; see the references at the end of the chapter.

## 14.1 Elliptic functions: generalities

In this section we define elliptic functions and deduce some basic properties. In the next section, we show that non-constant elliptic functions actually exist.

An *elliptic function* is a meromorphic function $f$, defined on all of $\mathbb{C}$, that is doubly periodic: there are non-zero constants $2\omega_1$, $2\omega_2$ in $\mathbb{C}$ such that $\omega_2/\omega_1$ is not real, and

$$f(u+2\omega_1) = f(u) = f(u+2\omega_2).$$

The numbers $2\omega_1$, $2\omega_2$ are the *periods* of $f$. (The reason for the factor 2 is that the half-periods $\omega_1$, $\omega_2$ play a significant role, see Propositions 14.1.4 and 14.1.5.)

The *period lattice* generated by $\omega_1$ and $\omega_2$ is the set

$$\Lambda = \Lambda(2\omega_1, 2\omega_2) = \{p : p = 2m\omega_1 + 2n\omega_2,\ m,n = 0, \pm 1, \pm 2, \dots\}.$$

Each $p \in \Lambda$ is also a period: $f(u+p) = f(u)$. Note that other choices of periods can generate the same lattice, and therefore the same class of elliptic functions, for example,

$$\omega_1' = \omega_1 + \omega_2, \quad \omega_2' = \omega_2 \tag{14.1.1}$$



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_14

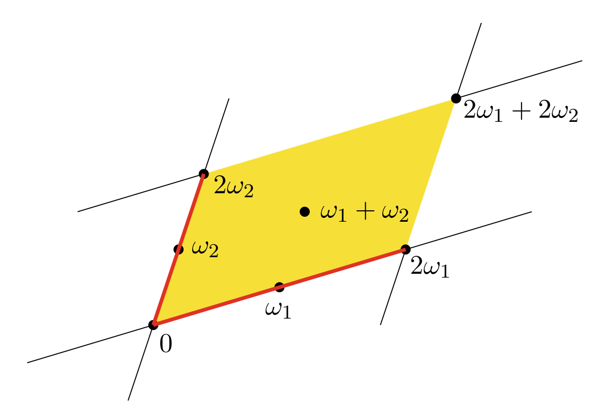



**Fig. 14.1** Period parallelogram

or

$$\omega_1' = -\omega_2, \quad \omega_2' = \omega_1 + \omega_2. \tag{14.1.2}$$

Usually it is assumed that the generators $2\omega_j$ are numbered so that

$$\operatorname{Im} \frac{\omega_2}{\omega_1} > 0. \tag{14.1.3}$$

The changes (14.1.1) and (14.1.2) preserve the condition (14.1.3).

An elliptic function $f$ is completely known once we know it on any *period parallelogram* $\Pi_a$:

$$\Pi_a = \Pi_a(2\omega_1, 2\omega_2) = \{u : u = a + 2s\omega_1 + 2t\omega_2,\ 0 \le s,t < 1\};$$

see Figure 14.1 for $\Pi_0$. In what follows we shall often tacitly choose $a$ so that $f$ has no zeros or poles on the boundary $\partial \Pi_a$ of $\Pi_a$. We write $\Pi$ for $\Pi_0$. The collection

$$\{\Pi_p : p \in \Lambda\} \tag{14.1.4}$$

is a covering of $\mathbb{C}$ by disjoint sets.

**Proposition 14.1.1.** *The set of elliptic functions with a given period lattice is closed under addition, multiplication, and division (by a function that is not identically zero). Each non-constant elliptic function has poles and is determined up to a multiplicative constant by its zeros and poles, counted according to multiplicity.*

*Proof:* The set of meromorphic functions is closed under these operations, and periodicity is clearly preserved. If two such functions $f$ and $g$ have the same zeros and poles, then the quotient $f/g$ is an entire function. Periodicity implies that $f/g$ has the same values on each set of the cover (14.1.4), so $f/g$ is bounded. By Liouville's theorem, Theorem 1.2.7, $f/g$ is constant. □

**Remark.** The first appearance of Liouville's theorem seems to have been in Liouville's treatment of elliptic functions. He used it as it was used here, to show that an entire elliptic function is constant.

**Proposition 14.1.2.** *A non-constant elliptic function $f$ has at least two poles, counting multiplicity, in each period parallelogram. Moreover, $f$ takes each (finite or infinite) value the same number of times in each period parallelogram.*

*Proof:* As noted above, if $f$ is not constant, then it cannot be entire. Therefore $f$ has at least one pole in each $\Pi_a$. The sum of the residues, counting multiplicity, is

$$\frac{1}{2\pi i}\int_\Gamma f(u)\,du,$$

where $\Gamma$ is the boundary $\partial\Pi_a$, oriented in the usual way. Because of periodicity, integrals over opposite sides cancel, so the integral is zero. Therefore either there is a multiple pole, or else several simple poles, the sum of whose residues is zero.

The integral

$$\frac{1}{2\pi i}\int_\Gamma \frac{f'(u)}{f(u)}\,du$$

counts the number of zeros minus the number of poles in the parallelogram. This integral is also zero, so the number of zeros equals the number of poles. This argument, applied to $f(u)-b$, shows that $f$ takes the value $b$ the same number of times, as well. (This number is called the *order* of $f$.) This number is independent of the base point $a$ of the period parallelogram, since it depends continuously on $a$, but is an integer. □

Another integration over the boundary of $\Pi_a$ shows that not only are the zeros and poles equal in number, but there is a constraint on how they are positioned.

**Proposition 14.1.3.** *Suppose that $f$ is an elliptic function whose zeros and poles in the period parallelogram $\Pi$, counting multiplicity, are $a_1,\dots,a_k$ and $b_1,\dots,b_k$, respectively. Then*

$$p \;=\; (a_1+a_2+\cdots+a_k)-(b_1+b_2+\cdots+b_k) \tag{14.1.5}$$

*belongs to $\Lambda$.*

*Proof:* Assume first that there are no zeros or poles on the boundary $\Gamma$ of $\Pi$. The right side of (14.1.5) is equal to

$$\frac{1}{2\pi i}\int_\Gamma z\frac{f'(z)}{f(z)}\,dz = \frac{1}{2\pi i}\left[-2\omega_2\int_0^{2\omega_1}\frac{f'(z)}{f(z)}\,dz+2\omega_1\int_0^{2\omega_2}\frac{f'(z)}{f(z)}\,dz\right].$$

The integrals on the right are equal to $i$ times the change in the argument of $f(z)$ as $z$ goes from 0 to $2\omega_1$ or $2\omega_2$. Since $f$ returns to its initial value at each of these points, the change in the argument is an integral multiple $m_j$ of $2\pi$. Therefore

$$p = 2m_1\omega_1 - 2m_2\omega_2 \in \Lambda. \tag{14.1.6}$$

If there are zeros or poles on the boundary, we may shift slightly, replacing the endpoints by $\varepsilon, 2\omega_j + \varepsilon$. □

According to Proposition 14.1.2, the simplest possibility for a non-constant elliptic function $f$ is that it has order two: either two simple poles whose residues have opposite signs, or one double pole. We could add to the simplicity by asking that $f$ be symmetric about its pole(s): odd, in the case of simple poles, or even, in the case of a double pole. We may assume, also, that the double pole, or one of the simple poles, is at the origin.

**Proposition 14.1.4.** *Suppose that $f$ is an odd elliptic function of order 2 with periods $2\omega_1$, $2\omega_2$, and a pole at the origin. Then the zeros and poles of $f$ are simple. The zeros and the other poles in the period parallelogram $\Pi$ are located at the half-periods $\{\omega_1, \omega_2, \omega_3\}$, where $\omega_3 = \omega_1 + \omega_2$. Moreover, $f$ is odd around each of the points $\omega_j$.*

*Proof:* We start with the elementary observation that if $g$ is odd and has period $2\omega$, then $g$ is also odd around $\omega$:

$$g(\omega - u) = -g(\omega + u).$$

See Exercise 3. It follows that $f$ is odd around each of the points $\omega_1$, $\omega_2$, $\omega_3$. Therefore each is a zero or pole of $f$; see Exercise 4. Since $f$ is assumed to have order 2, it has two poles and two zeros, counting multiplicity, so the zeros and the other poles take up the three half-periods $\omega_j$. □

**Remark.** Given the distribution of the zeros and other poles among the half-periods $\omega_j$, the function $f$ in Proposition 14.1.4 can be normalized by choosing the residue at the origin or the derivative at one of the zeros.

For the second possibility, an even elliptic function of order 2 with a double pole at the origin, one can always add a constant. The location of the two zeros (or double zero) in $\Pi$ depends on that constant. The additive and multiplicative constants can be fixed by the condition

$$f(u) = \frac{1}{u^2} + O(u^2) \tag{14.1.7}$$

near the origin. The best approach seems to be to consider the derivative, which would be odd, of order 3, with a triple pole at the origin, and with

$$f'(u) = -\frac{2}{u^3} + O(u) \tag{14.1.8}$$

near the origin.

**Proposition 14.1.5.** *Suppose that $g$ is an odd elliptic function of order 3 with periods $2\omega_1$, $2\omega_2$ and a triple pole at the origin. Then $g$ has three simple zeros in $\Pi$ at*

*the half-periods $\{\omega_j\}$, and is uniquely determined by the condition (14.1.8). Furthermore, there is a unique even elliptic function $f$ such that $f' = g$ and $f$ satisfies the condition (14.1.7).*

*Proof:* As in the previous proof, each of the half-periods must be a zero or pole of $g$. Since $g$ is assumed to have order 3, each of the $\omega_j$ must be a simple zero. The residue of $g$ at the origin is zero: see the proof of Proposition 14.1.2. Therefore the condition (14.1.8) determines $g$ uniquely.

The associated function $f$ must be given by

$$f(u) \;=\; \frac{1}{u^2} + \int_0^u h(t)\,dt, \qquad h(u) \;=\; g(u) + \frac{2}{u^3}.$$

By assumption, $h(u) = O(u)$ near the origin, so $f$ satisfies condition (14.1.7). Since $f' = g$ is odd, $f$ is even. The translates $f(u + 2\omega_j)$ have the same derivative as $f$, so they differ from $f$ by constants $c_j$. Note that $h$ is odd, so

$$c_j \;=\; f(\omega_j) - f(-\omega_j) \;=\; \int_{-\omega_j}^{\omega_j} h(t)\,dt \;=\; 0.$$

Therefore $f$ has periods $2\omega_j$. □

**A look ahead.** In the case of Jacobi elliptic functions, the periods are expressed in terms of real constants $K$ and $K'$. There are three basic functions of order 2, $\mathrm{sn}$, $\mathrm{cn}$, and $\mathrm{dn}$, each of which has simple poles and simple zeros. The periods of $\mathrm{sn}$ are $4K$ and $2iK'$, of $\mathrm{cn}$ are $4K$ and $2K + 2iK'$, and of $\mathrm{dn}$ are $2K$ and $4iK'$. (Thus each can be considered as having order 4, with periods $4K$, $4iK'$.)

The basic elliptic function in the Weierstrass theory, denoted $\wp$, is even and characterized by the conditions

$$\wp(u) \;=\; \frac{1}{u^2} + O(1), \quad u \in \Pi.$$

It is worth emphasizing that, to this point, we have not shown that non-constant elliptic functions actually exist. The existence statement in Proposition 14.1.5 was based on the assumption of the existence of the function $g$ with the stated properties.

## 14.2 Theta functions

As noted earlier, the usual normalization of the periods is to number them so that $\omega_2/\omega_1 = \tau$ has positive imaginary part. Replacing $f$ with $g(u) \;=\; f(2\omega_1 u)$, we may normalize further to periods $1, \tau$, with $\tau \in \mathbb{C}_+$.

The basic idea of the construction of elliptic functions via theta functions is that although a non-constant elliptic function cannot be entire, it may be the quotient of

two entire functions that are as close as possible to being doubly periodic. To this end we look for an entire function

$$\Theta(u) \;=\; \Theta(u|\tau)$$

such that

$$\Theta(u+1) \;=\; \Theta(u), \qquad \Theta(u+\tau) \;=\; a(u)\,\Theta(u), \tag{14.2.1}$$

where the function $a$ is as simple as possible. The conditions imposed on $\Theta$ imply that $a$ should be entire, nowhere zero, and have period 1.

Another desirable condition is that $\Theta$ have a single zero in a period parallelogram. If the boundary is $\Gamma$, this amounts to asking that

$$1 \;=\; \frac{1}{2\pi i}\int_\Gamma \frac{\Theta'(u)}{\Theta(u)}\,du. \tag{14.2.2}$$

Now $\Theta'/\Theta$ has period 1, so the two integrals over sides not parallel to the real axis cancel. On the upper boundary $\Theta'/\Theta$ differs from its value on the lower boundary by $a'/a$. Therefore (14.2.2) reduces to

$$1 \;=\; -\frac{1}{2\pi i}\int_0^1 \frac{a'(u)}{a(u)}\,du,$$

and the simplest solution is $a'/a \equiv -2\pi i$, or $a = ce^{-2\pi i u}$, where $c$ is a non-zero constant. The standard choice is $c = -1$, so

$$a(u) \;=\; -e^{-2\pi i u}, \tag{14.2.3}$$

which is indeed entire, periodic, and never zero. Thus our conditions on $\Theta$ are

$$\Theta(u+1) \;=\; \Theta(u), \qquad \Theta(u+\tau) \;=\; -e^{-2\pi i u}\Theta(u).$$

To determine $\Theta$ we begin with a formal Fourier expansion

$$\Theta(u) \;=\; \sum_{-\infty}^{\infty} a_n\,e^{2n\pi i u} \;=\; \sum_{-\infty}^{\infty} a_n\,p(u)^{2n}, \qquad p(u) \equiv e^{i\pi u}. \tag{14.2.4}$$

Then $p(u+\tau) = p(u)q$ with $q = e^{i\pi\tau}$, so (14.2.1) and (14.2.4) yield the formal identity

$$\Theta(u+\tau) \;=\; \sum_{-\infty}^{\infty} a_n\,p^{2n}\,q^{2n} \;=\; -p^{-2}\sum_{-\infty}^{\infty} a_n\,p^{2n} \;=\; -\sum_{-\infty}^{\infty} a_n\,p^{2n-2}.$$

Equating coefficients of $p^{2n-2}$ gives

$$a_n \;=\; -q^{2(n-1)}a_{n-1}.$$

Taking $a_0 = 1$, we find for positive $n$ that

$$a_n = (-1)^n q^{n(n-1)}. \tag{14.2.5}$$

For negative indices we start with

$$q^{-2n} a_{-n} = -a_{1-n}$$

and obtain

$$a_{-n} = (-1)^n q^{n(n+1)} = (-1)^n q^{(-n)(-n-1)},$$

which confirms (14.2.5) for every index. Thus we have a formal expression for $\Theta$:

$$\Theta(u) = \sum_{n=-\infty}^{\infty} (-1)^n p(u)^{2n} q^{n(n-1)}, \qquad p(u) = e^{i\pi u}, \quad q = e^{i\pi\tau}. \tag{14.2.6}$$

**Proposition 14.2.1.** *Suppose $\tau \in \mathbb{C}_+$. The series (14.2.6) converges and defines an entire function with the properties*

$$\Theta(u+1) = \Theta(u), \qquad \Theta(u+\tau) = -e^{-2\pi i u}\Theta(u). \tag{14.2.7}$$

*The zeros of $\Theta$ are the points of the period lattice*

$$\Lambda(1,\tau) = \{m + n\tau : m, n = 0, \pm 1, \pm 2, \dots\}.$$

*Proof:* Convergence follows from the fact that $\mathrm{Re}\,(i\tau) < 0$, so $|q| < 1$. The coefficients have moduli that are $O(|p(u)^{2n} q^{n^2}|)$. This decreases very rapidly as $|n| \to \infty$, uniformly for $u$ in bounded sets. Therefore (14.2.6) defines an entire function. Periodicity follows from the periodicity of $p$. The identity for $\Theta(u+\tau)$ follows from (14.2.5). By construction there is a unique zero in each period parallelogram. But

$$\begin{aligned}
\Theta(0) &= \sum_{n=-\infty}^{\infty} (-1)^n q^{n(n-1)} \\
&= \sum_{n=1}^{\infty} (-1)^n q^{n(n-1)} + \sum_{m=0}^{\infty} (-1)^m q^{(m+1)m} \\
&= \sum_{n=1}^{\infty} \left[ (-1)^n q^{n(n-1)} + (-1)^{n-1} q^{n(n-1)} \right] = 0.
\end{aligned}$$

It follows from this and the properties (14.2.7) that the zeros of $\Theta$ are precisely the points of $\Lambda(1,\tau)$. □

Propositions 14.1.4 and 14.1.5 indicate the special importance of the half-periods, which in this case are $\frac{1}{2}$ and $\frac{1}{2}\tau$.

**Proposition 14.2.2.** *The function $\Theta$ has the following properties:*

$$
\begin{aligned}
&p(u)^{-1}\Theta(u) \quad \textit{is an odd function;}\\
&p(u)^{-1}\Theta(u\pm\tfrac{1}{2}) \quad \textit{is an even function;}\\
&\Theta(u+\tfrac{1}{2}\tau) \quad \textit{is an even function;}\\
&\Theta(u+\tfrac{1}{2}\tau\pm\tfrac{1}{2}) \quad \textit{is an even function;}\\
&p(u)^{-2}q\,\Theta(u-\tfrac{1}{2}\tau) \;=\; -\Theta(u+\tfrac{1}{2}\tau);\\
&p(u)^{-2}q\,\Theta(u-\tfrac{1}{2}\tau\pm\tfrac{1}{2}) \;=\; \Theta(u+\tfrac{1}{2}\tau\pm\tfrac{1}{2}).
\end{aligned}
$$

The proof is left as Exercise 8.

The functions described here are close to the quartet of Jacobi theta functions:

$$\theta_1(u) = i\sum_{-\infty}^{\infty}(-1)^n p^{2n-1} q^{(n-\frac{1}{2})^2}; \tag{14.2.8}$$

$$\theta_2(u) = \sum_{-\infty}^{\infty} p^{2n-1} q^{(n-\frac{1}{2})^2}; \tag{14.2.9}$$

$$\theta_3(u) = \sum_{-\infty}^{\infty} p^{2n} q^{n^2}; \tag{14.2.10}$$

$$\theta_4(u) = \sum_{-\infty}^{\infty}(-1)^n p^{2n} q^{n^2}; \tag{14.2.11}$$

see Exercise 9.

## 14.3 Construction of elliptic functions

We begin by constructing the two types of elliptic function $f$ of order 2 discussed in Section 14.1. First, let us construct $f$ odd with simple zeros at 0 and $1/2$ and simple poles at $\tau/2$ and $(1+\tau)/2$. This suggests the quotient

$$g(u) \;=\; \frac{\Theta(u)\,\Theta(u-\frac{1}{2})}{\Theta(u-\frac{1}{2}\tau)\,\Theta(u-\frac{1}{2}-\frac{1}{2}\tau)}.$$

Now $a(u-c) = a(u)e^{2\pi i c}$, so

$$\frac{g(u+\tau)}{g(u)} \;=\; \frac{a(u)\,a(u-\frac{1}{2})}{a(u-\frac{1}{2}\tau)\,a(u-\frac{1}{2}-\frac{1}{2}\tau)} \;=\; \frac{e^{\pi i}}{e^{\pi i[\tau+(1+\tau)]}} \;=\; e^{-2\pi i\tau}.$$

It is easy enough to modify in order to get period $\tau$. Recall that each element of the period lattice $\Lambda$ is a zero of $\Theta$, so we can replace $\Theta(u-\frac{1}{2}\tau)$ by $\Theta(u+\frac{1}{2}\tau)$.

**Proposition 14.3.1.** *The function*

$$f(u) = \frac{\Theta(u)\Theta(u-\frac{1}{2})}{\Theta(u+\frac{1}{2}\tau)\Theta(u-\frac{1}{2}-\frac{1}{2}\tau)} \tag{14.3.1}$$

*is elliptic, with periods* $1, \tau$. *Moreover,* $f$ *has order 2 and is odd.*

*Proof:* The function $f$ is clearly meromorphic and has period 1. A calculation as above shows that $f$ also has period $\tau$. Since $\Theta$ is entire and $u = 0$ is its only zero in the period parallelogram $\Pi$, it follows that the zeros of $f$ in $\Pi$ are simple zeros at $u = 0$ and $u = \frac{1}{2}$. Similarly, the poles of $f$ in $\Pi$ are simple poles at $u = \frac{1}{2}\tau$ and $u = \frac{1}{2}(1+\tau)$. Therefore $f$ is elliptic of order 2. The fact that $f$ is odd follows from Proposition 14.2.2; see Exercise 10. □

For a second way to construct this type of function, see Exercise 12.

To construct a function of the type in Proposition 14.3.2, it is enough to construct (a multiple of) the derivative.

**Proposition 14.3.2.** *Suppose* $\tau \in \mathbb{C}_+$. *The function*

$$g(u) = \frac{\Theta(u+\frac{1}{2})\,\Theta(u+\frac{1}{2}\tau)\,\Theta(u-\frac{1}{2}-\frac{1}{2}\tau)}{\Theta(u)^3} \tag{14.3.2}$$

*is odd, elliptic of order 3 with periods* $1, \tau$ *and has a triple pole at the origin.*

*Proof:* The periodicity argument is essentially the same as the argument for Proposition 14.3.1. The fact that $g$ is odd follows from Proposition 14.2.2; see Exercise 11. □

The preceding two propositions are examples of a completely general construction of elliptic functions.

**Theorem 14.3.3.** *Suppose* $\tau \in \mathbb{C}_+$. *Every non-constant elliptic function* $F$ *with periods* $1, \tau$ *has the form*

$$F(u) = c\,\frac{\Theta(u-a_1)\,\Theta(u-a_2)\cdots\Theta(u-a_k)}{\Theta(u-b_1)\,\Theta(u-b_2)\cdots\Theta(u-b_k)}, \tag{14.3.3}$$

*where* $c$ *is constant,* $k \geq 2$ *and*

$$a_1 + a_2 + \cdots + a_k = b_1 + b_2 + \cdots + b_k. \tag{14.3.4}$$

*Conversely, every function of this form is elliptic of order* $k$.

*Proof:* Given an elliptic function $f$ of order $k$, let the $\{a_j\}$ and $\{b_j\}$ be as in Proposition 14.1.3, and let $p$ be given by (14.1.5). Since $p$ is a period of $f$, we may replace

$a_1$ by $a_1 - p$ and accomplish (14.3.4) without changing the zeros of $f$. Let $F$ be given by (14.3.3) with $c = 1$. The condition (14.3.4) is sufficient for $F$ to have period $\tau$ in addition to period 1. Thus $F$ is elliptic and has the same zeros and poles as $f$, so $f/F$ is constant.

Conversely, each $F$ constructed in this way is elliptic of order k. □

## 14.4 Integrating elliptic functions

Let $Z$ be the logarithmic derivative of $\Theta$:

$$Z(u) = \frac{\Theta'(u)}{\Theta(u)}. \tag{14.4.1}$$

This has period 1 and a simple pole at each lattice point. The second property in (14.2.1) implies that

$$Z(u+\tau) = Z(u) - 2\pi i. \tag{14.4.2}$$

Therefore any linear combination of translates of $Z$ whose coefficients sum to zero,

$$\sum_{j=1}^{n} c_j Z(u - b_j), \qquad \sum_{j=1}^{n} c_j = 0, \tag{14.4.3}$$

has period $\tau$ and is elliptic. Since $Z$ is a derivative, combinations (14.4.3) can be integrated, in terms of linear combinations of logarithms of translates of $\Theta$.

The identity (14.4.2) implies that the derivative

$$Z'(u) = \frac{\Theta''(u)}{\Theta(u)} - Z(u)^2 \tag{14.4.4}$$

has period 1 and a double pole at each lattice point.

Suppose now that $f$ is an arbitrary non-constant elliptic function with periods 1 and $\tau$ that we wish to integrate. Suppose first that $f$ has only simple poles. As noted in the proof of Proposition 14.1.2, the sum of the residues in a period parallelogram is zero. Therefore $f$ differs from a combination (14.4.3) by a constant function.

If $f$ has a pole of order $k > 1$, at $u = b$, then there is a constant $c$ such that

$$f(u) - cZ^{(k-1)}(u-b)$$

has a pole of order $< k$ (or a removable singularity) at $b$. Continuing, we can express $f$ in the form

$$f(u) = \left[\sum c_m Z^{(m)}(u - b_m)\right]' + g(u),$$

where $g$ has only simple poles or is constant.

## Exercises

1. Suppose that $\operatorname{Im}(\omega_2/\omega_1) > 0$, and

$$\omega_1' = a\omega_1 + b\omega_2; \quad \omega_2' = c\omega_1 + d\omega_2,$$

where $a, b, c, d$ are complex constants. Show that the pairs $\omega_1, \omega_2$ and $\omega_1', \omega_2'$ generate the same period lattice if and only if $a, b, c, d$ are integers and $ad - bc = \pm 1$. Show that $\operatorname{Im}(\omega_2'/\omega_1') > 0$ if and only if $ad - bc = 1$.

2. Show that if $f$ is an elliptic function of order $m$, its even and odd parts

$$f_e(u) = \frac{1}{2}[f(u) + f(-u)], \qquad f_o(u) = \frac{1}{2}[f(u) - f(-u)]$$

are elliptic functions with the same periods, and orders $\leq m$.

3. Suppose that $f$ has period $2\omega$.

   (a) Suppose that $f$ is odd. Show that $f$ is odd around $\omega$.

   (b) Suppose that $f$ is even. Show that $f$ is even around $\omega$.

4. Suppose that $f$ is elliptic with periods $2\omega_j$ and is odd. Show that each of the points 0, $\omega_1$, $\omega_2$, and $\omega_1 + \omega_2$ is a zero or pole of $f$.
5. Suppose that $f$ is an elliptic function of order $m > 0$. Show that $1/f$ is also elliptic of order $m$.
6. Suppose $f$ has order $m$. Show that the order of $f'$ can be anything from $m+1$ to $2m$.
7. Suppose that $f$ and $g$ are elliptic functions with the same periods. Suppose that $f$ has order $m > 0$ and $g$ has order $n \geq m$. What can be said about (a) the order of the sum $f + g$, (b) the order of the product $fg$, and (c) the order of the quotient $f/g$?
8. Use the definition of $\Theta$ and the properties

$$p(u + \tfrac{1}{2})^2 = -p(u)^2, \qquad p(u + \tfrac{1}{2}\tau)^2 = p(u)^2 q$$

to prove Proposition 14.2.2.

9. Write each of the Jacobi theta functions (14.2.8)–(14.2.11) as a multiple of one of the functions in Proposition 14.2.2.
10. Use Proposition 14.2.2 to verify that (14.3.1) defines an odd function.
11. Use Proposition 14.2.2 to verify that (14.3.2) defines an odd function.
12. Show that the quotient

$$g(u) = \frac{\Theta(u)}{p(u)\Theta(u + \frac{1}{2}\tau)}$$

has the same properties as the function in Proposition 14.3.1, but with respect to periods 2 and $\tau$ rather than 1 and $\tau$.

13. Suppose that $f$ is even, elliptic of order 2 with a double pole in the period parallelogram at $u = 0$, and satisfies

$$f(u) \;=\; \frac{1}{u^2} + O(u^2)$$

as $u \to 0$. Show that $(f')^2 - 4f^3$ is elliptic of order at most two and satisfies an equation

$$(f')^2 \;=\; 4f^3 + af + b$$

for some constants $a$ and $b$.

14. Suppose that $f$ is elliptic of order two with simple poles at $u = 0$ and $u = \omega_2$ in the period parallelogram, and is odd. Normalizing, we may assume that in the period parallelogram,

$$f(u) \;=\; \frac{1}{u} - \frac{1}{u-\omega_2} + O(1).$$

Show that $(f')^2 - 4f^4$ is even and periodic, with order 2, and

$$(f')^2 \;=\; 4f^4 + af^2 + b$$

for some constants $a$ and $b$.

15. Suppose that $f$ is as in Exercise 13 and $g$ is an elliptic function with the same periods. Show that $g$ can be written as a linear combination of functions of the form

$$P(f(u-a)), \qquad f'(u-b)\,Q(f(u-b)), \qquad h(u-c),$$

where $P$ and $Q$ are polynomials and $h$ has only simple poles or is constant.

16. This is the first in a series of exercises leading to a proof of the *Jacobi triple product formula*, one version of which is

$$\Theta(u+\tfrac{1}{2}) \;=\; \prod_{n=1}^{\infty}(1-q^{2n})[1-p(u)^2q^{2n-1}][1-p(u)^{-2}q^{2n-1}]. \qquad (14.4.5)$$

Deduce from the location of the zeros of the left side that

$$\Theta(z+\tfrac{1}{2}\tau) \;=\; c(z,\tau)\prod_{n=1}^{\infty}[1-p(z)^2q^{2n-1}][1-p(z)^{-2}q^{2n-1}], \qquad (14.4.6)$$

where $c(z,\tau)$ is an entire function of $z$ with no zeros.

17. Compute the changes in the product in (14.4.6) if $z$ is changed to $z+1$ or to $z+\tau$, and show that $c(z,\tau)$ is independent of $z$: thus $c(z,\tau) = G(\tau)$.
18. Show that the function $G$ in Exercise 17 has limit 1 as $\operatorname{Im}\tau \to \infty$.
19. Show that

$$\begin{aligned}\Theta\left(\frac{1}{4}+\frac{1}{2}\tau\right) &= G(\tau)\prod_{n=1}^{\infty}(1+q^{4n-2}) \\ &= \sum_{-\infty}^{\infty}(-i)^n q^{n^2} \;=\; \sum_{-\infty}^{\infty}(-1)^m q^{(2m)^2}.\end{aligned}$$

20. Use Exercise 19 to show that

$$\frac{G(\tau)}{G(4\tau)} = \prod_{n=1}^{\infty} \frac{(1-q^{8n-4})^2}{1+q^{4n-2}}$$
$$= \prod_{n=1}^{\infty}(1-q^{8n-4})(1-q^{4n-2}).$$

21. Show that for $|z| < 1$,

$$\prod_{n=1}^{\infty}(1-z^{2n-1}) = \prod_{n=1}^{\infty}\frac{(1-z^n)}{1-z^{2n}}.$$

22. Use Exercises 20 and 21 to show that

$$\frac{G(\tau)}{G(4\tau)} = \frac{\prod_{n=1}^{\infty}(1-q^{2n})}{\prod_{n=1}^{\infty}(1-q^{8n})}$$

and therefore

$$\frac{G(\tau)}{G(4^m\tau)} = \frac{\prod_{n=1}^{\infty}(1-q^{2n})}{\prod_{n=1}^{\infty}(1-q^{4^m 2n})}.$$

23. Use Exercises 18 and 22 to prove the triple product formula (14.4.5).
24. Note that (14.4.5) is the same as

$$\prod_{n=1}^{\infty}(1-q^{2n})[1-p(u)^2q^{2n-1}][1-p(u)^{-2}q^{2n-1}] = \sum_{-\infty}^{\infty}(-1)^n p(u)^{2n}q^{n^2}. \tag{14.4.7}$$

25. Deduce other versions of the triple product, and write the forms that correspond to the form (14.4.7):

$$\Theta(u) = \prod_{n=1}^{\infty}(1-q^{2n})(1-p^2q^{2n-2})(1-p^{-2}q^{2n});$$
$$\Theta(u+\tfrac{1}{2}) = \prod_{n=1}^{\infty}(1-q^{2n})(1+p^2q^{2n-2})(1+p^{-2}q^{2n});$$
$$\Theta(u+\tfrac{1}{2}+\tfrac{1}{2}\tau) = \prod_{n=1}^{\infty}(1-q^{2n})(1+p^2q^{2n-1})(1+p^{-2}q^{2n-1}).$$

## Remarks and further reading

There is a vast literature on elliptic functions. Weil [139] presents an approach due to Eisenstein. Modern references include Siegel [128], Walker [138], and Armitage and Eberlein [10]. For the early history, see Roy [121].

For more on the history and modern developments of theta functions, see Kempf [76] and Murty [101]. For more discussion of the literature, see the remarks at the end of Chapter 14 of [16]. For an efficient presentation of Riemann's theory of theta functions in several variables, with applications, see Dubrovin [38].

# Chapter 15
# Jacobi elliptic functions




Jacobi elliptic functions are a realization of one of the simplest cases of elliptic functions, as described in Chapter 14: functions with two simple poles in a period parallelogram that are odd around each pole. These functions come up naturally in certain problems of mechanics, such as the motion of an ideal pendulum. In pure mathematics they arise, for example, in connection with maps from the upper half plane to a parallelogram. In this chapter we begin with the pendulum equation and derive the properties of the functions associated to it. The triple of Jacobi functions $\mathrm{sn}$, $\mathrm{cn}$, $\mathrm{dn}$ is closely analogous to the pair of trigonometric functions sine and cosine, and satisfy similar identities.

## 15.1 The pendulum equation

An ideal pendulum is probably the simplest equation of one-dimensional mechanics that leads to functions beyond the standard fare of calculus. Consider the motion of such a pendulum: the weight moves along an arc of a circle, its position at time $t$ marked by an angle $\theta(t)$ from the (downward) vertical; see Figure 15.1. Therefore acceleration is proportional to $\theta''$. The force, gravity, is essentially constant, directed downward, so a little geometry shows that the component in the direction of the pendulum's motion is proportional to $\sin\theta$. Thus (after scaling) Newton's equation of motion has the form

$$2\theta'' = -\sin\theta. \tag{15.1.1}$$

Multiplying both sides of (15.1.1) by $\theta'$ and integrating give

$$(\theta')^2 = \cos\theta - a. \tag{15.1.2}$$

We assume that $0 < a < 1$, which has the effect of insuring that (15.1.2) has a non-zero solution, and that the angle $\theta$ stays strictly between the upward verticals $-\pi/2$ and $\pi/2$. Let $w(t) = \sin(\frac{1}{2}\theta(t))$, so that (15.1.2) takes the form




R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_15

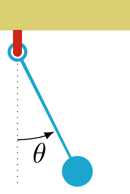



**Fig. 15.1** Pendulum

$$(w')^2 \;=\; A(1-k^{-2}w^2)(1-w^2),$$

with $0<k<1$. We let $w(t)=ku(t)$ and rescale time to get an equation that can be written as

$$\frac{dt}{du} \;=\; \frac{1}{\sqrt{(1-u^2)(1-k^2u^2)}}. \tag{15.1.3}$$

We expect the same position $u$ to occur at different times $t$, i.e. $t$ is not a single-valued function of $u$. In any event we are more interested in the inverse function: $u(t)$ as a function of $t$. This function is defined implicitly by (15.1.3)

$$t \;=\; \int_{u(0)}^{u(t)} \frac{ds}{\sqrt{(1-s^2)(1-k^2s^2)}}. \tag{15.1.4}$$

Integrating (15.1.3) leads to the function

$$F(z) \;=\; F(k,z) \;=\; \int_0^z \frac{ds}{\sqrt{(1-s^2)(1-k^2s^2)}}. \tag{15.1.5}$$

In the next section we discuss the properties of this function.

## 15.2 Properties of the map F

We begin by looking at $F$ as a function on the upper half plane $\mathbb{C}_+$.

**Theorem 15.2.1.** *The function $F$ maps the upper half plane $\mathbb{C}_+$ to a rectangle $R$.*

*Proof:* Taking into account the fact that the integral is odd, $F$ maps the interval $[-1,1]$ onto the interval $[-K,K]$, where

$$K \;=\; K(k) \;=\; \int_0^1 \frac{ds}{\sqrt{(1-s^2)(1-k^2s^2)}}. \tag{15.2.1}$$

The interval $[1,1/k]$ is mapped onto the interval from $K$ to $K+iK'$, where

$$K' = K'(k) = \int_1^{1/k} \frac{ds}{\sqrt{(s^2-1)(1-k^2s^2)}}. \tag{15.2.2}$$

The integral from $1/k$ to $\infty$ along the real axis is real and increasing. Since the integrand is even, the integral from $-\infty$ to $-1/k$ is real and decreasing, and has the same value. Therefore the image of $[1/k,\infty] \cup [-\infty,-1/k]$ is the interval from $K+iK'$ to $-K+iK'$. By symmetry, the image of $[-1/k,-1]$ is the interval from $-K+iK'$ to $-K$, and the image of $[-1,0]$ is $[-K,0]$. To show that the upper half plane $\mathbb{C}_+$ maps to the interior of the rectangle, it is enough to note that $F(i)$ has positive imaginary part. $\square$

Let $P(z) = (1-z^2)(1-k^2z^2)$. The function $\sqrt{P(z)}$ has a well-defined branch, on the complement of the two real intervals $[-1/k,-1]$, $[1,1/k]$, that is positive for $z$ in the positive imaginary axis. Of course $\sqrt{P(z)}$ has a second branch on this same complement that is negative on the positive imaginary axis. We can think of $\sqrt{P(z)}$ as a single-valued function if we take as its domain two copies of the complex plane, (an "upper sheet" and a "lower sheet," respectively) properly joined across the slits. Better yet, consider two copies of the Riemann sphere, with the slits opened up and glued together in such a way that the result is, topologically, a torus—see Figure 7.3. This is also the natural domain for the reciprocal $1/\sqrt{P(z)}$.

We extend $F$ by taking the integral (15.1.5) over other paths in the two-sheeted representation of the domain of the integrand. By Cauchy's theorem, the integral is not changed if we deform a portion of the path that lies entirely in one of the two sheets without leaving that sheet, i.e. without crossing one of the slits.

Consider a path from $0 = 0_+$ in the upper sheet back to $0_+$ that passes once through each slit: first the slit on the right, then the left; see Figure 15.2. The path may be taken along intervals 0 to 1 in the upper sheet, 1 to $-1$ in the lower sheet, and $-1$ to 0 in the upper sheet. The integral along the path is $4K$. The integral in the opposite direction is $-4K$. This path can be retraced multiple times in one direction or the other, and combined with any given path from $0_+$ to $z$. Therefore $F(z)$ is only defined up to addition of integer multiples of $4K$.

Consider next a path from $0_+$ to itself in the upper sheet that goes around the slit on the right. We take it to proceed on line segments in the upper sheet from 0 to 1, from 1 to $1/k$ along the upper margin of the slit, from $1/k$ back to 1 along

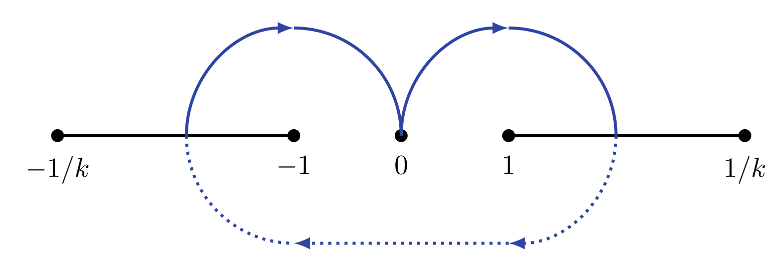



**Fig. 15.2** Path from zero to zero

the lower margin (thus with the integrand having opposite sign), and from 1 back to 0. The contributions between 0 and 1 cancel, and the value of the integral is $2iK'$. Combining this with the previous argument, we conclude that $F(z)$ is defined only up to terms $4mK + 2inK'$, $m, n$ integers.

Consider a path from $z_+$ to $z_-$, the representative of $z = z_+$ on the lower sheet, that crosses to the lower sheet at $z = 1$ and proceeds first to $0_-$ and then to $z_-$. The result is the identity

$$F(z_-) = 2K - F(z_+). \tag{15.2.3}$$

## 15.3 The Jacobi functions

As in the case of the revised pendulum equation (15.1.3), we look for a function sn that inverts the function $F$. In view of the properties of $F$, we must have

$$\mathrm{sn}\,(z+4K) = \mathrm{sn}\,(z+2iK') = \mathrm{sn}\,z. \tag{15.3.1}$$

The desired function can be defined implicitly by

$$z = \int_0^{\mathrm{sn}\,z} \frac{ds}{\sqrt{(1-s^2)(1-k^2s^2)}}, \tag{15.3.2}$$

with $z$ and 0 in the upper sheet, along any path in the domain of definition of the integrand. Each path can be continuously deformed to one that includes a certain number of copies of the two paths from 0 to 0 considered above; this ambiguity is reflected in (15.3.1), which implies more generally that $\mathrm{sn}\,(z+p) = \mathrm{sn}\,(z)$ for each $p$ in the period lattice

$$\Lambda = \Lambda(k) = \{4mK + 2niK' \, : \, m, n \in \mathbb{Z}\}. \tag{15.3.3}$$

We shall see that sn is meromorphic, so sn is an elliptic function. Assuming this for the moment, it follows that sn takes each finite or infinite value the same number of times in each period rectangle

$$\Pi_a = \{u : u = a + 4sK + 2tiK', \, 0 \le s, t < 1\}. \tag{15.3.4}$$

We know that sn has simple zeros at 0 and $2K$ only, so it takes each value twice. Moreover sn is odd:

$$\mathrm{sn}\,(-z) = -\,\mathrm{sn}\,z. \tag{15.3.5}$$

Therefore sn is odd around each half-period in $\Pi_0$, i.e. $2K$, $iK'$, and $2K + iK'$, so each of these points is either a zero or a simple pole. It follows that $iK'$ and $2K + iK'$ must be simple poles.

Conversely, if we show that there are simple poles at these points and no other in $\Pi_o$, then periodicity implies that sn is everywhere meromorphic and therefore an elliptic function. The points in $\Pi_0$ where sn is infinite are the values of

$$\int_0^\infty \frac{ds}{\sqrt{(1-s^2)(1-k^2s^2)}}$$

over the positive imaginary axis, or first to 0 on the lower sheet and then over the positive imaginary axis. The change of variable $t = \sqrt{s^2-1}/\sqrt{1-k^2s^2}$ shows that the integral along the positive imaginary axis is

$$\begin{aligned}\int_0^{i\infty} \frac{ds}{\sqrt{(1-s^2)(1-k^2s^2)}} &= i\int_0^\infty \frac{dt}{\sqrt{(1+t^2)(1+k^2t^2)}}\\ &= i\int_1^{1/k} \frac{ds}{\sqrt{(s^2-1)(1-k^2s^2)}} \;=\; iK'.\end{aligned}$$

An examination of the integral shows that $iK'$ is a simple pole—see Exercise 3. Therefore $2K+iK'$ is a second simple pole and sn is indeed an elliptic function.

Since sn is odd, sn is odd around half-periods. Equation (15.2.3) implies that sn is even around the quarter-period $K$. Thus

$$\begin{aligned}\operatorname{sn}(2K-z) &= -\operatorname{sn}(2K+z);\\ \operatorname{sn}(iK'-z) &= -\operatorname{sn}(iK'+z);\\ \operatorname{sn}(2K+iK'-z) &= -\operatorname{sn}(2K+iK'+z);\\ \operatorname{sn}(K-z) &= \operatorname{sn}(K+z).\end{aligned} \tag{15.3.6}$$

From calculations above, and (15.3.6), we have specific values at certain combinations of half-periods and quarter-periods:

$$\begin{aligned}\operatorname{sn}0 = \operatorname{sn}(2K) &= 0;\\ \operatorname{sn}(iK') = -\operatorname{sn}(2K+iK') &= \infty; \qquad (15.3.7)\\ \operatorname{sn}K = -\operatorname{sn}(3K) &= 1;\\ \operatorname{sn}(K+iK') = -\operatorname{sn}(3K+iK') &= 1/k. \qquad (15.3.8)\end{aligned}$$

It follows from (15.3.2) that the derivative

$$\operatorname{sn}' = \sqrt{(1-\operatorname{sn}^2)(1-k^2\operatorname{sn}^2)}. \tag{15.3.9}$$

For $z$ near 0 we may define functions $\operatorname{cn} z = \operatorname{cn}(k,z)$ and $\operatorname{dn} z = \operatorname{dn}(k,z)$ by taking the principal branches of the square root:

$$\operatorname{cn} = \sqrt{1-\operatorname{sn}^2}, \qquad \operatorname{dn} = \sqrt{1-k^2\operatorname{sn}^2}. \tag{15.3.10}$$

Each of the functions $1-\operatorname{sn}^2$, $1-k^2\operatorname{sn}^2$ has two double zeros in $\Pi_0$, so this branch of the square root can be continued analytically throughout $\Pi_0$, apart from the poles of sn. Since sn is even or odd about each half-period, the same is true of cn and dn. Neither function vanishes at 0 or $2K$, so they are even around these points. Similarly cn is even around $2K+iK'$ and dn is even around $K$. Now cn has simple zeros at $K$ and $3K$, and dn has simple zeros at $K+iK'$ and $3K+iK'$. Therefore cn and dn

are odd at these points, respectively. Both functions have simple poles at $iK'$ and $2K+iK'$, so they are odd at these points as well. It follows that these functions are also doubly periodic, but with different periods:

$$\mathrm{cn}\,u = \mathrm{cn}\,(u+4K) \;=\; \mathrm{cn}\,(u+2K+2iK'); \tag{15.3.11}$$

$$\mathrm{dn}\,u = \mathrm{dn}\,(u+2K) \;=\; \mathrm{dn}\,(u+4iK'). \tag{15.3.12}$$

See Exercise 5.

It follows from these definitions and analytic continuation that cn and dn are even functions. By definition

$$\mathrm{sn}^2+\mathrm{cn}^2 \;=\; 1, \qquad \mathrm{dn}^2+k^2\mathrm{sn}^2 \;=\; 1, \qquad \mathrm{dn}^2-k^2\mathrm{cn}^2=1-k^2. \tag{15.3.13}$$

The derivatives satisfy

$$\mathrm{sn}' \;=\; \mathrm{cn}\,\mathrm{dn}\,, \qquad \mathrm{cn}' \;=\; -\,\mathrm{sn}\,\mathrm{dn}\,, \qquad \mathrm{dn}' \;=\; -k^2\,\mathrm{sn}\,\mathrm{cn}\,. \tag{15.3.14}$$

The functions sn, cn, dn are the *Jacobi elliptic functions*. From the definitions, together with (15.3.7) and (15.3.8), one can compute the remaining values in the following table:

| | $0$ | $K$ | $2K$ | $3K$ | $iK'$ | $K+iK'$ | $2K+iK'$ | $3K+iK'$ |
|---|---|---|---|---|---|---|---|---|
| sn | $0$ | $1$ | $0$ | $-1$ | $\infty$ | $k^{-1}$ | $\infty$ | $-k^{-1}$ |
| cn | $1$ | $0$ | $-1$ | $0$ | $\infty$ | $-ik'k^{-1}$ | $\infty$ | $ik'k^{-1}$ |
| dn | $1$ | $k'$ | $1$ | $k'$ | $\infty$ | $0$ | $\infty$ | $0$ |

The triple of functions sn, cn, dn is clearly analogous to the pair of functions sine and cosine, solutions of the equation $(S')^2=1-S^2$. There are analogues for these Jacobi functions of the addition theorems for sine and cosine. Here we outline a derivation of the addition theorems, with the details left to the exercises.

Consider the function

$$F(z) \;=\; \mathrm{cn}\,z\,\mathrm{cn}\,(z-u)+A\,\mathrm{sn}\,z\,\mathrm{sn}\,(z-u), \qquad u \neq iK'+2mK+2niK', \tag{15.3.15}$$

where $A$ is constant. This is an elliptic function with periods $2K$ and $2iK'$; Exercise 18. The constant $A$ can be chosen so that $F$ does not have a pole at $z=iK'$. Therefore $F$ has no poles and, by Proposition 14.1.1, is constant. The constant can be computed by taking $z=0$. Observe that $F(0)=\mathrm{cn}\,u$. For $z$ close to 0,

$$\begin{aligned} F(z) &= F(0)+F'(0)z+O(z^2) \\ &= \mathrm{cn}\,(u)+\big[\mathrm{cn}'(0)\,\mathrm{cn}\,(-u)+\mathrm{cn}\,(0)\,\mathrm{cn}'(-u)\big]\,z \\ &\quad +A\,\mathrm{sn}'(0)\,\mathrm{sn}\,(-u)\,z+O(z^2). \end{aligned} \tag{15.3.16}$$

The values $\mathrm{cn}'(0)$ and $\mathrm{dn}'(0)$ can be obtained from (15.3.9) and (15.3.14), yielding

$$F(z) \;=\; \mathrm{cn}\,u+[\mathrm{sn}\,u\,\mathrm{dn}\,u-A\,\mathrm{sn}\,u]\,z+O(z^2). \tag{15.3.17}$$

Therefore $A = \text{dn}\, u$ and $F(z) = \text{cn}\, u$, which gives

$$\text{dn}\, u\, \text{sn}\, z\, \text{sn}\,(z-u) + \text{cn}\, z\, \text{cn}\,(z-u) \;=\; \text{cn}\, u. \tag{15.3.18}$$

A similar argument starting with

$$G(z) \;=\; \text{dn}\, z\, \text{dn}\,(z-u) + A\, \text{sn}\, z\, \text{sn}\,(z-u) \tag{15.3.19}$$

leads to the equation

$$k^2\, \text{cn}\, u\, \text{sn}\, z\, \text{sn}\,(z-u) + \text{dn}\, z\, \text{dn}\,(z-u) \;=\; \text{dn}\, u.. \tag{15.3.20}$$

Setting $u = z + w$ in (15.3.18) and 15.3.20) gives

$$-\,\text{dn}\,(z+w)\, \text{sn}\, z\, \text{sn}\, w + \text{cn}\, z\, \text{cn}\, w \;=\; \text{cn}\,(z+w); \tag{15.3.21}$$
$$-k^2\, \text{cn}\,(z+w)\, \text{sn}\, z\, \text{sn}\, w + \text{dn}\, z\, \text{dn}\, w \;=\; \text{dn}\,(z+w). \tag{15.3.22}$$

Setting $u = -w$ in (15.3.18) and (15.3.20) gives

$$\text{dn}\, w\, \text{sn}\, z\, \text{sn}\,(z+w) + \text{cn}\, z\, \text{cn}\,(z+w) \;=\; \text{cn}\, w; \tag{15.3.23}$$
$$k^2\, \text{cn}\, w\, \text{sn}\, z\, \text{sn}\,(z+w) + \text{dn}\, z\, \text{dn}\,(z+w) \;=\; \text{dn}\, w. \tag{15.3.24}$$

Equations (15.3.21), (15.3.22), and (15.3.23) are linear in $\text{sn}\,(z+w)$, $\text{cn}\,(z+w)$, and $\text{dn}\,(z+w)$. Solving for these expressions gives the *addition formulas*.

**Theorem 15.3.1.**

$$\text{sn}\,(z+w) \;=\; \frac{\text{sn}\, z\, \text{cn}\, w\, \text{dn}\, w + \text{sn}\, w\, \text{cn}\, z\, \text{dn}\, z}{1 - k^2\, \text{sn}^2 z\, \text{sn}^2 w};$$

$$\text{cn}\,(z+w) \;=\; \frac{\text{cn}\, z\, \text{cn}\, w - \text{sn}\, z\, \text{dn}\, z\, \text{sn}\, w\, \text{dn}\, w}{1 - k^2\, \text{sn}^2 z\, \text{sn}^2 w};$$

$$\text{dn}\,(z+w) \;=\; \frac{\text{dn}\, z\, \text{dn}\, w - k^2\, \text{sn}\, z\, \text{cn}\, z\, \text{sn}\, w\, \text{cn}\, w}{1 - k^2\, \text{sn}^2 z\, \text{sn}^2 w}.$$

## 15.4 Elliptic curves: Jacobi parametrization

This section is parallel to, but independent of, Section 16.3. An algebraic curve in $\mathbb{C}^2$ is a set

$$C_P \;=\; \{(w,z) \in \mathbb{C}^2 : P(w,z) \;=\; 0\},$$

where $P$ is a polynomial in two variables. The simplest non-trivial example involves $P$ of degree two, not the product of two linear terms. With a linear change of variables $P$ may be put in the form $P(w,z) \;=\; w^2 + z^2 = 1$, i.e.

$$w^2 \;=\; 1 - z^2.$$

The corresponding set $C_P = \{(w,z) : z^2 = 1-w^2\}$ can be parametrized by the map

$$u \to (\cos u, \sin u) = (f'(u), f(u)), \quad f(u) = \sin u.$$

In view of equations (15.3.10) and (15.3.14) the curve defined by analogous equation

$$w^2 = Q(z), \tag{15.4.1}$$

where $Q$ is quartic (degree 4), can be parametrized by the map

$$u \to (\mathrm{sn}'(u), \mathrm{sn}(u)).$$

As shown in Section 16.3, the curve defined by equation (15.4.1) when $Q$ is cubic can also be parametrized by an elliptic function and its derivative. Thus, for $Q$ of degree 2, 3, or 4, the curve associated to equation (16.3.1) can be parameterized by functions meromorphic in the entire plane. Picard showed that this is no longer possible as soon as $Q$ has degree greater than 4; see Theorem 9.4.2.

## Exercises

1. Suppose $w_1 < w_2 < w_3 < w_4$ are four distinct real numbers. Show that there is a linear fractional transformation, taking $\mathbb{C}_+$ to itself, that maps these points, respectively, to $-1/k < -1 < 1 < 1/k$ for some real $k$.
2. Prove (15.3.6).
3. Show that sn has simple poles at $iK'$ and $-2K+iK'$ with residues $1/k$ and $-1/k$, respectively.
4. Find analogues of (15.3.9) for $\mathrm{cn}'$ and $\mathrm{dn}'$.
5. Use the symmetry properties of cn and dn to verify (15.3.11) and (15.3.12).
6. Show that none of $2K$, $2iK'$, or $K+iK'$ is a period of cn.
7. Show that neither $K$ nor $2iK'$ is a period of dn.
8. Verify the values of cn and dn in the table.
9. Show that the residues of cn are $-i/k$ at $iK'$ and $i/k$ at $2K+iK'$.
10. Show that the residues of dn are $-i$ at $iK'$ and $i$ at $2K+iK'$.
11. Suppose that $f$ has period $4K$ and is either even or odd around $u = 2K$. Prove that $g(u) = f(u)f(u-z)$ has period $2K$. Deduce that the function $F$ in (15.3.15) is elliptic with the stated periods.
12. Show that the constant $A$ in (15.3.15) can be chosen so that $F$ is holomorphic at $iK'$, hence constant.
13. Use (15.3.10) and (15.3.14) to justify the passage from (15.3.16) to (15.3.17).
14. Verify (15.3.18).
15. Carry out the argument leading from (15.3.19) to (15.3.20).
16. Verify the formulas in Theorem 15.3.1.
17. Show that the formula for $\mathrm{sn}(z+w)$ can be expressed entirely in terms of $\mathrm{sn}\, z$, $\mathrm{sn}\, w$, $\mathrm{sn}'z$, $\mathrm{sn}'w$, and find analogous expressions for $\mathrm{cn}(z+w)$ and $\mathrm{dn}(z+w)$.

18. (a) Show that sn and cn are odd around $u = K$, while $dn$ is even around $u = K$.
(b) Use part (a) to show that if (15.3.15) has no pole at $iK'$, then it has no pole at $2K + iK'$.

## Remarks and further reading

See the references for Chapter 14.

# Chapter 16
# Weierstrass elliptic functions




This chapter depends on Chapter 14 but not on Chapter 15. Here we look for a more direct approach to elliptic functions with given periods:

$$f(z+2\omega_1) = f(z) = f(z+2\omega_2), \qquad \operatorname{Im}\frac{\omega_2}{\omega_1} > 0. \tag{16.0.1}$$

The associated period lattice is

$$\Lambda = \Lambda(2\omega_1, 2\omega_2) = \{2m\,\omega_1 + 2n\,\omega_2 : n, m = 0, \pm 1, \pm 2, \dots\} \tag{16.0.2}$$

and the associated period parallelograms are

$$\Pi_a = \Pi_a(2\omega_1, 2\omega_2) = \{a + 2s\omega_1 + 2t\omega_2, : 0 \le s, t < 1\}, \quad a \in \mathbb{C}. \tag{16.0.3}$$

Weierstrass's approach to elliptic functions was to start with the period lattice and construct associated functions explicitly.

## 16.1 The Weierstrass $\wp$ function

The simplest elliptic function of order two that is canonically associated to the period lattice $\Lambda$ of (16.0.2) is one that is even and has a double pole at each point $p$ of $\Lambda$, with

$$f(u) = \frac{1}{u^2} + O(u^2)$$

near the origin. Weierstrass constructed such a function $\wp$ explicitly. It is tempting to use the recipe

$$\sum_{p\in\Lambda} \frac{1}{(z-p)^2}$$

for $\wp$, since this (formal) sum appears to be doubly periodic, with a double pole at each point $p \in \Lambda$. However the series does not converge.



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_16

We begin, instead, with a proposal for the derivative:

$$\wp'(u) = \sum_{p\in\Lambda} \frac{2}{(p-u)^3} = \sum_{m,n=-\infty}^{\infty} \frac{2}{(2m\omega_1+2n\omega_2-u)^3}.$$

This converges for each $u$ not in $\Lambda$. In fact, given $u \in \mathbb{C}\setminus\Lambda$, the number of lattice points $p$ in the annulus $r \le |p-u| < 2r$ is bounded by a multiple of the area $3\pi r^2$, and each point in the annulus contributes $O(1/r^3)$ to the sum. Thus the sum of the moduli of such terms is dominated by a fixed constant times $r^2/r^3 = 1/r$. Taking $r = 2^n$, $n = 0,1,2,\ldots$, we find that the sum of the moduli of the terms of the series with $|p-u| \ge 1$ is bounded by a fixed constant. It is easily seen that the series converges uniformly on each compact subset of the complement of $\Lambda$ and defines an odd, periodic, meromorphic function with order 3. Moreover

$$\wp'(u) = -\frac{2}{u^3} + O(u) \tag{16.1.1}$$

as $u \to 0$. Therefore we proceed as in the proof of Proposition 14.1.5 to define the *Weierstrass* $\wp$ function by

$$\wp(u) = \wp(2\omega_1, 2\omega_2; u) = \frac{1}{u^2} + \int_0^u \left\{\wp'(s) + \frac{2}{s^3}\right\} ds$$
$$= \frac{1}{u^2} + \int_0^u \sum_{p\in\Lambda,\, p\neq 0} \frac{2}{(p-s)^3}\, ds.$$

The series in the integrand converges, since, for fixed $u$ not in $\Lambda$, the terms are $O(p^{-3})$ for large $p$. Therefore we may integrate term-by-term and find that

$$\wp(u) = \frac{1}{u^2} + \sum_{p\in\Lambda,\, p\neq 0} \left[\frac{1}{(u-p)^2} - \frac{1}{p^2}\right]. \tag{16.1.2}$$

Notice that the integral is independent of the path (which is assumed to avoid $\Lambda$), since integrating around a point of $\Lambda$ picks up the residue at that pole, which is zero. As noted in the proof of Proposition 14.1.5, the function defined in this way is even and periodic, with periods $2\omega_j$.

Returning to $\wp'$, since it is odd, with a pole of order 3 at the origin, the three half-periods $\omega_1$, $\omega_2$, $\omega_3 = \omega_1 + \omega_2$ are simple zeros of $\wp'$. It follows that $\wp$ takes each of the corresponding values $e_j = \wp(\omega_j)$ with multiplicity two. Since $\wp$ has order 2, the values $e_j$ must be distinct.

**Proposition 16.1.1.** *The Weierstrass function $\wp$ satisfies a differential equation of the form*

$$(\wp')^2 = 4\wp^3 - g_2\wp - g_3. \tag{16.1.3}$$

*Proof:* The functions $\wp^3$ and $(\wp')^2$ have $\Lambda$ as period lattice, and each has a pole of order 6 at the origin. Since $\wp'$ is odd, the singular part of $(\wp')^2$ at the origin has

no terms of order less than two. In view of (16.1.1), the singular part of $\wp^3$ at the origin, which is even, has no term of lower order other than order 2. Checking the terms of order at most 6, we find that there are constants $g_2$, $g_3$ such that the doubly periodic function

$$(\wp')^2 - 4\wp^3 + g_2\wp + g_3$$

is entire and vanishes at $u = 0$. □

The equation (16.1.3) can be written

$$\wp' = \sqrt{Q(\wp)}, \qquad Q(t) = 4t^3 - g_2 t - g_3.$$

Inverting, we are led to the integral equation

$$u = u_0 + \int_{\wp(u_0)}^{\wp(u)} \frac{ds}{\sqrt{4s^3 - g_2 s - g_3}}. \tag{16.1.4}$$

Equations (16.1.3) and (16.1.1), and the fact that the $e_j$ are distinct, imply that

$$Q(t) \equiv 4t^3 - g_2 t - g_3 = 4(t - e_1)(t - e_2)(t - e_3), \quad e_j = \wp(\omega_j). \tag{16.1.5}$$

Therefore

$$\begin{aligned} e_1 + e_2 + e_3 &= 0; \\ 4(e_1e_2 + e_2e_3 + e_3e_1) &= -g_2; \\ 4e_1e_2e_3 &= g_3. \end{aligned}$$

We can characterize the elliptic functions with period lattice $\Lambda$.

**Theorem 16.1.2.** *Each elliptic function with period lattice $\Lambda$ is a rational function of translates of $\wp$ and $\wp'$.*

*Proof:* Let $f$ be such a function, and assume first that $f$ is even. Subtracting some linear combination of powers of $\wp$, we may assume that $f$ is regular and non-zero at the origin. The assumption that $f$ is even means that the zeros and poles in the period parallelogram

$$\Pi = \{2s\omega_1 + 2t\omega_2 : -1 \le s, t < 1\}$$

come in pairs $c, c^*$ symmetric about the midpoint $\omega_3$ of $\Pi$: $\omega_3 - c = \omega_3 + c^*$. Take one representative $a_j$ from each pair of zeros and one representative $b_j$ from each pair of poles, $j = 1, 2, \ldots, n$. Then

$$g(u) = \prod_{k=1}^{n} \frac{\wp(u) - \wp(a_k)}{\wp(u) - \wp(b_k)}$$

has the same zeros and poles as $f$, counting multiplicity, so the quotient $f/g$ is constant.

Suppose now that $f$ is odd. Then $f = g\wp'$, where $g = f/\wp'$ is even. □

## 16.2 Integration of elliptic functions

What happens if we integrate $\wp$? The *Weierstrass zeta function* $\zeta(u)$ is determined uniquely by the conditions

$$\zeta'(u) \;=\; -\wp(u), \qquad \zeta(-u) \;=\; -\zeta(u).$$

Note that $\wp(u) - 1/u^2$ is regular at the origin. For $u$ not in the period lattice $\Lambda$ we take

$$\zeta(u) \;=\; \frac{1}{u} + \int_0^u \left[\frac{1}{s^2} - \wp(s)\right] ds.$$

The residues of the integrand at the non-zero points of the period lattice $\Lambda$ all vanish, so the integration can be taken along any path that avoids $\Lambda$, and $\zeta$ is single-valued and meromorphic. The derivative is $-\wp$. Expanding the integrand,

$$\zeta(u) \;=\; \frac{1}{u} + \int_0^u \sum_{p\in\Lambda,\, p\neq 0} \left[\frac{1}{p^2} - \frac{1}{(s-p)^2}\right] ds.$$

Again, the terms in the series are $O(p^{-3})$ so the term-by-term integration is justified and

$$\zeta(u) \;=\; \frac{1}{u} + \sum_{p\in\Lambda,\, p\neq 0} \left[\frac{u}{p^2} + \frac{1}{u-p} + \frac{1}{p}\right]. \tag{16.2.1}$$

Thus $\zeta(u) \;=\; 1/u + O(1)$ near the origin. This is the only pole in $\Pi$, so $\zeta$ cannot have $\Lambda$ as period lattice. For $u + \omega_j$ not in $\Lambda$, since $\wp$ has period $2\omega_j$, the integral

$$\int_u^{u+2\omega_j} \wp(s)\, ds \;=\; \zeta(u+2\omega_j) - \zeta(u)$$

exists and is constant. Taking $u = -\omega_j$, we have, for $u$ not in $\Lambda$,

$$\zeta(u+2\omega_j) \;=\; \zeta(u) + 2\zeta(\omega_j). \tag{16.2.2}$$

(The sum in (16.2.1) converges, since it is obtained by integrating an absolutely convergent sum, or one can check directly that the term indexed by $p$ is $O(|p|^{-3})$.)

Consider now a general elliptic function $f$. If $f$ has simple poles at $b_1, \ldots, b_n$ in $\Pi_0$ with residues $a_1, \ldots, a_k$, then

$$g(u) \;=\; f(u) - \sum_{j=1}^{k} a_j\, \zeta(u - b_j)$$

has only poles of higher order in $\Pi_0$. Moreover we know from integrating $f$ around the boundary of $\Pi_a$, some $a \approx 0$, that $\sum_{j=1}^{k} a_j = 0$. In view of this and (16.2.2),

$g$ is elliptic. Since $g$ has only multiple poles, it is, up to an additive constant, a linear combination of translates of derivatives of $\zeta' = -\wp$. Thus $f$ itself is the sum of a linear combination of translates of $\zeta$, translate of $\wp$ and its derivatives, and a constant.

We have reduced the question of integrating an elliptic function to the question of integrating (translates of) $\zeta$. This leads to the *Weierstrass sigma function*, characterized by the conditions

$$\frac{\sigma'}{\sigma} = \zeta, \quad \lim_{u\to 0} \frac{\sigma(u)}{u} = 1.$$

Then $\log\sigma$ is an integral of $\zeta$. We take

$$\log\sigma(u) = \log u + \int_0^u \left[\zeta(s) - \frac{1}{s}\right] ds.$$

Once again we may put in the series expansion of the integrand and integrate term-by-term to obtain

$$\log\sigma(u) = \log u + \sum_{p\in\Lambda,\, p\neq 0} \left[\log\left(1-\frac{u}{p}\right) + \frac{u}{p} + \frac{u^2}{2p^2}\right].$$

It follows that

$$\sigma(u) = u \prod_{p\in\Lambda,\, p\neq 0} \left(1-\frac{u}{p}\right) \exp\left(\frac{u}{p} + \frac{u^2}{2p^2}\right) \tag{16.2.3}$$

is an entire function whose zeros are the lattice points $\Lambda$. The function $\sigma$ is a close analogue of the function $\Theta$ of Chapter 14. In fact

$$\frac{d}{du}\left\{\log\frac{\sigma(u+2\omega_j)}{\sigma(u)}\right\} = \zeta(u+2\omega_j) - \zeta(u) = 2\eta_j$$

so

$$\log\frac{\sigma(u+2\omega_j)}{\sigma(u)} = 2\eta_j u + C_j, \qquad C_j \text{ constant.}$$

In particular, taking $u = -\omega_j$ we have

$$\log\frac{\sigma(\omega_j)}{\sigma(-\omega_j)} = -2\eta_j\omega_j + C_j.$$

Since $\sigma$ is odd, the left side is (some determination of) $\log(-1)$, so

$$C_j = \log(-1) + 2\eta_j\omega_j$$

and

$$\log\frac{\sigma(u+2\omega_j)}{\sigma(u)} = \log(-1) + 2\eta_j\omega_j + 2\eta_j u.$$

Thus

$$\sigma(u+2\omega_j) = -e^{2\eta_j(\omega_j+u)}\sigma(u). \tag{16.2.4}$$

This identity leads to a representation of elliptic functions as quotients of entire functions; see Exercise 9.

**Theorem 16.2.1.** *Each elliptic function $f$ of order $m$ with periods $2\omega_1$, $2\omega_2$ can be written in the form*

$$f(u) = c\,\frac{\prod_{j=1}^{m}\sigma(u-a_j)}{\prod_{j=1}^{m}\sigma(u-b_j)},$$

*where $c$ is constant.*

If $n$ is an integer $\geq 2$, the sum

$$G_n = \sum_{p\in\Lambda, p\neq 0}\frac{1}{p^{2n}} \tag{16.2.5}$$

is called the *Eisenstein series of order $n$*. In particular $\wp''(z)-6z^{-4}$ has a removable singularity at $z=0$ with value $6G_2$. Other relations of the $G_n$ with $\wp$, with each other and with the coefficients $g_2$, $g_3$ in the differential equation (16.1.3) are explored in Exercises 4–8.

Why only even powers in (16.2.5)? For $n \geq 1$ the series

$$\sum_{p\in\Lambda,\, p\neq 0}\frac{1}{p^{2n+1}}$$

converges absolutely, but each $0 \neq p \in \Lambda$ can be matched with $-p$, so the sum is zero.

## 16.3 Elliptic curves: Weierstrass parametrization

This section is parallel to, but independent of, Section 15.4. An algebraic curve in $\mathbb{C}^2$ is the set

$$C_P = \{(w,z)\in\mathbb{C}^2 : P(w,z) = 0\},$$

where $P$ is a polynomial in two variables. The simplest non-trivial example involves $P$ of degree two, not the product of two linear terms. With a linear change of variables $P$ may be put in the form $P(w,z) = w^2+z^2 = 1$, i.e.

$$w^2 = 1-z^2.$$

The corresponding set $C_P = \{(w,z) : z^2 = 1-w^2\}$ can be parametrized by the map

$$u \to (\cos u, \sin u) = (f'(u), f(u)), \quad f(u) = \sin u.$$

In view of equation (16.1.3), the curve is defined by analogous equation

$$w^2 = Q(z), \tag{16.3.1}$$

where $Q$ is cubic, which can be parametrized by the map

$$u \to (\wp'(u), \wp(u)). \tag{16.3.2}$$

(To be precise, the cubic $Q$ in (16.3.1) can easily be normalized to have the form $4z^3 - g_2 z - g_3$. If we assume that $Q$ has three distinct roots, then it does arise in the equation of a suitable $\wp$, and the map (16.3.2) is surjective: Exercise 11.)

As shown in Section 15.4, the curve defined by equation (16.3.1) when $Q$ is quartic (degree 4) can also be parametrized by an elliptic function and its derivative. Thus for $Q$ of degree 2, 3, or 4, the curve associated to equation (16.3.1) can be parameterized by functions meromorphic in the entire plane. Picard showed that this is no longer possible as soon as $Q$ has degree $> 4$: Theorem 9.4.2.

## 16.4 Addition on the curve

The basic periodic functions in one variable satisfy “addition theorems,” for example,

$$\sin u + \sin v = 2 \sin\left(\frac{u+v}{2}\right) \cos\left(\frac{u-v}{2}\right).$$

The Weierstrass function $\wp$ satisfies an equation of this type, based on the following observation: each complex line in $\mathbb{C}^2$ of the form

$$L = L_{a,b} = \{(z,w) : w = az + b\}, \quad a \neq 0,$$

meets the curve $C$ in $\mathbb{C}^2$, defined by the equation

$$C = \{(z,w) : w^2 = 4z^3 - g_2 z - g_3\}, \tag{16.4.1}$$

in three points, counting multiplicity. In fact substituting $az + b$ for $w$ reduces the equation to a cubic equation for $z$. If $(z,w)$ is such a solution then

$$0 = (w + az + b)(w - az - b) = w^2 - (az+b)^2 \tag{16.4.2}$$

$$= 4z^3 - a^2 z^2 - (2ab + g_2)z - (b^2 + g_3). \tag{16.4.3}$$

Therefore if the three points of intersections of $L$ and $C$ are $(z_j, w_j)$, then the $z_j$ are the roots of the cubic in (16.4.3), so

$$4(z_1 + z_2 + z_3) = a^2. \tag{16.4.4}$$

On the other hand, the two equations $w_j = az_j + b$, $j = 1, 2$ determine

$$a = \frac{w_2 - w_1}{z_2 - z_1}, \tag{16.4.5}$$

and combining this with (16.4.4) gives

$$z_3 = -z_1 - z_2 + \frac{1}{4}\left(\frac{w_2 - w_1}{z_2 - z_1}\right)^2. \tag{16.4.6}$$

This leads to the *addition formula* for $\wp$.

**Theorem 16.4.1.** *If $\wp(u) \neq \wp(v)$, then*

$$\wp(u+v) = -\wp(u) - \wp(v) + \frac{1}{4}\left(\frac{\wp'(v) - \wp'(u)}{\wp(v) - \wp(u)}\right)^2. \tag{16.4.7}$$

*Proof:* The assumption that $\wp(u) \neq \wp(v)$ implies that there is a unique solution $(a,b) \in \mathbb{C}^2$ to the pair of equations

$$\wp'(u) = a\wp(u) + b, \qquad \wp'(v) = a\wp(v) + b.$$

The function

$$\wp'(z) - a\wp(z) - b$$

is elliptic of order 3 with a pole of order three at the origin. Therefore there are three zeros, counting multiplicity. The pole has residue zero, so the sum of the zeros is a lattice point, and we may assume that they are chosen so that the sum is zero. By the choice of $(a,b)$, two of the zeros are $u$ and $v$, so the third is $-(u+v)$. But $\wp$ is even, so $\wp(u+v) = \wp(-u-v)$. Therefore equation (16.4.6) gives (16.4.7). □

Theorem 16.4.1 can be interpreted as describing a geometric procedure that equips the elliptic curve $C$ with the structure of a commutative group. Note that if $R = (z,w)$ is a point of $C$, then so is $R^* = (z,-w)$. The group operation assigns to two points $P$, $Q$ on $C$ the point $R^*$, where $R$ is the third point of intersection of $C$ with the line that passes through $P$ and $Q$. (Figure 16.1 illustrates this; it shows the graph in $\mathbb{R}^2$ of $y^2 = P(x)$ for a choice of $P$ having three real zeros.)

This operation is clearly commutative, but it is not obvious, based on the geometric description alone, that it is associative. However, if $P = \wp(u)$, $Q = \wp(v)$, the preceding argument shows that $R$ can be taken to be $\wp(-u-v)$. Since $\wp$ is even and $\wp'$ is odd, it follows that $R^* = \wp(u+v)$. In other words, the parametrization of $C$ by the map (16.3.2) is a group homomorphism between $\mathbb{C}$, with addition as composition, and $C$ with the geometric composition as just described. (Note that we must extend both the geometric construction and the argument in the addition theorem to the case $P = Q$: see Exercise 12.)

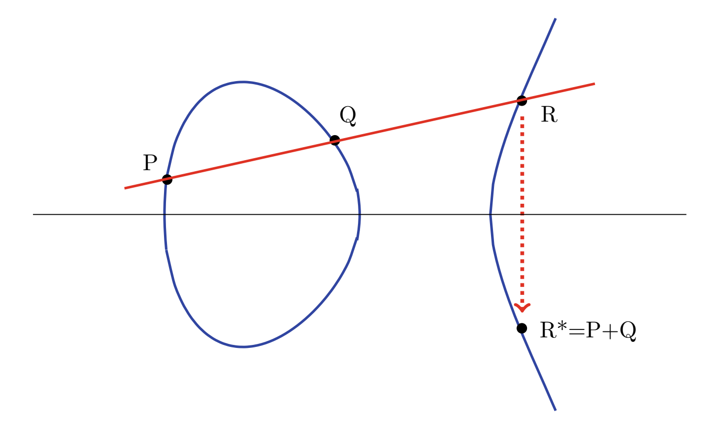



**Fig. 16.1** Addition on the curve

## Exercises

1. Find $\wp'''$ in terms of $\wp$ and $\wp'$.
2. Prove that
$$\wp''(u) = 6\wp(u)^2 - \tfrac{1}{2}g_2.$$
3. Prove that
$$\wp''(\omega_1) = 2(e_1 - e_2)(e_1 - e_3).$$
4. As noted above, for $n > 1$, $G_n$ is defined by
$$\sum_{p \in \Lambda,\, p \neq 0} \frac{1}{p^{2n}}.$$
Show that for $u$ close to the origin
$$\wp(u) = \frac{1}{u^2} + \sum_{n=1}^{\infty} (2n+1) G_{n+1} u^{2n}.$$
5. Use Exercise 4 to prove that the coefficients in the differential equation (16.1.3) are
$$g_2 = 60\,G_2, \quad g_3 = 140\,G_3.$$
6. Compute $G_2$ and $G_3$ as functions of the periods $2\omega_1, 2\omega_2$.
7. Use (16.1.3) and Exercise 5 to show that each $G_n$ is a polynomial in $g_2$ and $g_3$ with rational coefficients.
8. Show that for $n \geq 1$, $G_n$ has the Fourier expansion
$$G_n = 2\zeta(2n) + \frac{2(2\pi i)^{2n}}{(2n-1)\,!} \sum_{m=1}^{\infty} \sigma_{2n-1}(m) e^{2m\pi i}.$$

Here $\zeta$ is the Riemann zeta function (*not* the Weierstrass zeta function).

9. Use the analogue of Lemma 14.1.3 to prove Theorem 16.2.1; cf. the proof of Theorem 14.3.3.
10. Express $\wp$ and $\wp'$ as in Theorem 16.2.1.
11. Suppose that $Q(t) = 4t^3 - at - b$ has three distinct roots.

    (a) Show that there is a unique lattice $\Lambda$ such that the corresponding Weierstrass function $\wp$ satisfies $\wp' = Q(\wp)$.

    (b) Show that each point $(z, w)$ in the curve $w^2 = Q(t)$ is equal to $(\wp(u), \wp'(u))$ for some $u$ in the period parallelogram $\Pi$.

    (c) Show that in all cases of three distinct roots there is a rational transformation of the variable $t$ such that in the new variable the associated lattice is $\{m + in : m, n \in \mathbb{Z}\}$.
12. Extend the addition theorem to the case $u = v$ and express $\wp(2u)$ as a rational function of $\wp(u)$. How should the geometric construction of addition on the curve be interpreted in this case?

## Remarks and further reading

See the references for Chapter 14. For further applications of the Weierstrass theory, see Chapter 17.

# Chapter 17
# Automorphic functions and Picard's theorem




This chapter relies heavily on Chapter 16, with some reference to analytic continuation and conformal mapping, particularly Theorem 5.4.2.

Given a period lattice $\Lambda$ in $\mathbb{C}$, one wants to consider the meromorphic functions that have simple behavior with respect to affine transformations that map $\Lambda$ to itself. In the case of invariance under translation by the periods, these are the elliptic functions. In this chapter we introduce examples of an important class of functions with different invariance properties, the automorphic functions. In particular, the modular function $\lambda$ is used, as it was by Picard, to prove that an entire function cannot omit more than one value.

## 17.1 The elliptic modular function

We start by going back to the Weierstrass function $\wp$ and the associated zeros $e_j$ discussed earlier:

$$4\wp^3 - g_2\wp - g_3 \;=\; 4(\wp - e_1)(\wp - e_2)(\wp - e_3), \qquad e_j \;=\; \wp(\omega_j),$$

where $2\omega_1$ and $2\omega_2$ are the periods and $\omega_3 = \omega_2 + \omega_1$. We now consider the roots $e_j$ as functions of the periods. The *elliptic modular function* $\lambda$ associated with a given pair of periods is

$$\lambda \;=\; \frac{e_3 - e_2}{e_1 - e_2}. \tag{17.1.1}$$

It is clear from the series expansion of $\wp$, (16.1.2), that the roots $e_j$ are homogeneous functions of the periods, having degree $-2$:

$$e_j(2a\omega_1, 2a\omega_2) \;=\; a^{-2} e_j(2\omega_1, 2\omega_2), \qquad a \neq 0. \tag{17.1.2}$$



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_17

It follows that $\lambda$ is homogeneous of degree 0, and therefore depends only on the ratio $\tau = \omega_2/\omega_1$. Thus we consider $\lambda$ as a function of $\tau$. We shall always normalize by assuming $\operatorname{Im}\tau > 0$.

**Proposition 17.1.1.** *The modular function $\lambda(\tau)$ is holomorphic in $\mathbb{C}_+$ and does not take the values* 0 *or* 1.

*Proof.* The roots $e_j$ are holomorphic functions of $\omega_1$ and $\omega_2$ and depend only on the ratio $\omega_2/\omega_1$. The roots are distinct, so $\lambda$ is never 0 or 1. □

## 17.2 The modular group and the fundamental domain

The period lattice is mapped onto itself by certain linear transformations:

$$\begin{bmatrix}\omega_2'\\ \omega_1'\end{bmatrix} = \begin{bmatrix}a & b\\ c & d\end{bmatrix}\begin{bmatrix}\omega_2\\ \omega_1\end{bmatrix} = \begin{bmatrix}a\omega_2 + b\omega_1\\ c\omega_2 + d\,\omega_1\end{bmatrix}, \tag{17.2.1}$$

where $a, b, c, d$ are integers and the matrix is invertible. It follows that the determinant $ad - bc = \pm 1$. The corresponding action on the ratio $\tau$ is given by the linear fractional transformation

$$\tau' = \frac{a\tau + b}{c\tau + d}.$$

Since we insist that $\operatorname{Im}\tau' = \operatorname{Im}(\omega_2'/\omega_1')$ be positive we must have $ad - bc = 1$. The group $G$ of such matrices $g$ with integer entries is called the *modular group*. (We shall abuse the terminology here and identify each such matrix with the induced linear fractional transformation, and conversely.)

Let us try for a canonical choice of generators of the lattice $\Lambda = \{2m\omega_1 + 2n\omega_2\}$.

**Proposition 17.2.1.** *Generators $2\omega_j'$ can be chosen so that the ratio $\tau'$ belongs to the set*

$$\Delta = \{\tau' \in \mathbb{C}_+ : |\tau'| \ge 1,\ -\tfrac{1}{2} < \operatorname{Re}\tau' \le \tfrac{1}{2},\ \operatorname{Re}\tau' \ge 0 \text{ if } |\tau'| = 1\}. \tag{17.2.2}$$

(See Figure 17.1.)

*Proof:* Choose $2\omega_1' \in \Lambda$ with smallest modulus, and choose $\pm 2\omega_2' \in \Lambda$, not a multiple of $\omega_1'$, to have smallest modulus among such elements, with the sign chosen so that $\operatorname{Im}\tau' = \operatorname{Im}(\omega_2'/\omega_1')$ is positive. Then $|\tau'| \ge 1$. Moreover, by the choice of $\omega_2'$ we have

$$|\omega_2'| \le |\omega_2' \pm \omega_1'|,$$

which implies that $|\operatorname{Re}\tau'| \le \frac{1}{2}$. If $\operatorname{Re}\tau' = -\frac{1}{2}$, we may replace $\omega_2'$ with $\omega_2' + \omega_1'$. If $|\tau'| = 1$, we may replace the pair $\omega_1', \omega_2'$ with the pair $-\omega_2',\ \omega_1'$. (See Exercise 1.)
□

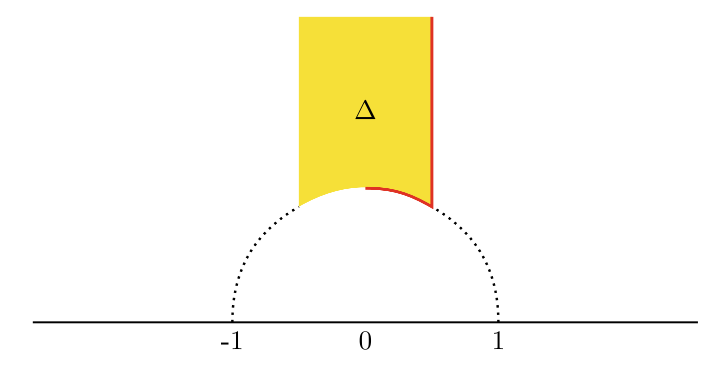



**Fig. 17.1** The region $\Delta$

Next, we look at the behavior of $\lambda$ under the action of the modular group on the periods. The function $\wp$ is determined by the lattice, so it is invariant under this action. Therefore each $g \in G$ permutes the set of zeros $\{e_j\}$. Note that if any of $a,b,c,d$ in (17.2.1) is changed by adding or subtracting an even integer, then the values of the $e'_j = \wp(\omega'_j)$ are unchanged. Therefore $G$ can be replaced, in its actions on the $e_j$ and on $\lambda$, by its reduction mod 2.

The modular group is generated by the two elements

$$R = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \qquad T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}. \tag{17.2.3}$$

The proof of this fact is left as Exercise 2.

Let us consider the effect of $R$. We have

$$\omega'_2 = \omega_1, \qquad \omega'_1 = -\omega_2, \qquad \tau' = -1/\tau,$$

and it follows from this and periodicity that $e'_1 = e_2$, $e'_2 = e_1$, $e'_3 = e_3$. Therefore

$$\lambda\left(-\frac{1}{\tau}\right) = \frac{e_3 - e_1}{e_2 - e_1} = \frac{e_3 - e_2 + e_2 - e_1}{e_2 - e_1} = 1 - \lambda(\tau). \tag{17.2.4}$$

Similarly, let us consider the effect of $T$. Here

$$\omega'_2 = \omega_1 + \omega_2 = \omega_3, \quad \omega'_1 = \omega_1, \quad \omega'_3 = 2\omega_1 + \omega_2, \quad \tau' = \tau + 1.$$

Therefore $e'_1 = e_1$, $e'_2 = e_3$, $e'_3 = e_2$ and $\lambda(\tau+1) = (e_2 - e_3)/(e_1 - e_3)$. Thus

$$\frac{1}{\lambda(\tau+1)} = \frac{e_1 - e_3}{e_2 - e_3} = \frac{e_1 - e_2 + e_2 - e_3}{e_2 - e_3} = 1 - \frac{1}{\lambda(\tau)},$$

so

$$\lambda(\tau+1) = \frac{\lambda(\tau)}{\lambda(\tau)-1}. \tag{17.2.5}$$

Iterating,

$$\lambda(\tau+2) = \lambda(\tau). \tag{17.2.6}$$

It follows from (17.2.4) and (17.2.5) that

$$\lambda\left(\frac{1}{1-\tau}\right) = \frac{1}{1-\lambda(\tau)}; \tag{17.2.7}$$

$$\lambda\left(1-\frac{1}{\tau}\right) = 1-\frac{1}{\lambda(\tau)}. \tag{17.2.8}$$

It follows from (17.2.4) and (17.2.5), that if $g$ belongs to the modular group and $\lambda(\tau)$ is real, then $\lambda(g(\tau))$ is real.

**Proposition 17.2.2.** *For each $\tau \in \mathbb{C}_+$, there is an element $g$ in the modular group such that $\tau' = g(\tau)$ belongs to the set $\Delta$ of (17.2.2). Moreover, $\tau'$ is unique.*

*Proof:* The first assertion is just a restatement of Proposition 17.2.1. To prove uniqueness amounts to showing that if $\tau$ and $\tau'$ belong to $\Delta$, and they are related by a modular transformation

$$\tau' = \frac{a\tau+b}{c\tau+d}, \qquad a,b,c,d \in \mathbb{Z}, \quad ad-bc=1, \tag{17.2.9}$$

then $\tau' = \tau$. Note that (17.2.9) is equivalent to

$$\tau = \frac{d\tau'-b}{-c\tau'+a}. \tag{17.2.10}$$

Suppose that $c=0$. Then $ad=1$ and $\tau' = \tau \pm b$. If both $\tau'$ and $\tau$ belong to $\Delta$, then $|\operatorname{Re}\tau' - \operatorname{Re}\tau| < 1$, so $b=0$ and $\tau'=\tau$.

Suppose that $d=0$. Then $bc=-1$ and $\tau' = (-1/\tau) \pm a$. Since $-1/\tau = -\bar{\tau}/|\tau|^2$, the map $\tau \to -1/\tau$ takes $\Delta$ into the intersection of the unit disk with the strip where $-1/2 \le \operatorname{Re} z < 1/2$. If both $\tau$ and $\tau'$ belong to $\Delta$, we must have $|\operatorname{Re}(-1/\tau)| \le 1/2$. But also

$$|\operatorname{Re}\tau'| = \left|\operatorname{Re}\left(-\frac{1}{\tau}\right) \pm a\right| \le \frac{1}{2}.$$

Since $a$ is an integer, it follows that $a=0$ or $a=\pm 1$. A check of these three cases shows that the only possibilities are $|\tau| = |\tau'| = 1$ and $\operatorname{Re}\tau = -1/2$, $a=-1$ or $\operatorname{Re}\tau = 1/2$, $a=1$. In each case, $\tau'=\tau$.

The same argument, applied to (17.2.10), shows that $a=0$ implies $\tau=\tau'$.

Suppose, finally, that $acd \neq 0$. Note that

$$\operatorname{Im}\tau' = \frac{\operatorname{Im}\tau}{|c\tau+d|^2}, \qquad \operatorname{Im}\tau = \frac{\operatorname{Im}\tau'}{|c\tau'-a|^2}. \tag{17.2.11}$$

Again, $|\tau| \geq 1$, $|\operatorname{Re}\tau| \leq 1/2$, so

$$|c\tau+d|^2 \;=\; [c^2|\tau|^2+d^2]+2cd\operatorname{Re}\tau \;\geq\; 2|cd\tau|-2|cd\operatorname{Re}\tau| \;\geq\; |cd| \;\geq 1.$$

Similarly $|c\tau'-a|^2 \geq 1$. It follows from these inequalities that $\operatorname{Im}\tau = \operatorname{Im}\tau'$, $|\tau| = |\tau'| = 1$, and $|\operatorname{Re}\tau| = |\operatorname{Re}\tau'| = 1/2$. The last two identities reduce the possible locations of $\tau$ and $\tau'$ to two points, only one of which belongs to $\Delta$. Thus in all cases $\tau' = \tau$. $\square$

Combining Proposition 17.2.2 with (17.2.4) and (17.2.5), we see that $\lambda$ is completely determined by its action on the set (17.2.2), and this is not true of any proper subset of (17.2.2): $\Delta$ is a *fundamental domain* for the function $\lambda$. (This is standard terminology, so we make an exception to our practice of taking a "domain" to be an open set.)

## 17.3 A closer look at $\lambda$; Picard's theorem

Let us look more closely at the numerator and denominator of $\lambda$. For this purpose, we may take the periods to be 1 and $\tau$, $\tau \in \mathbb{C}_+$. Then

$$\begin{aligned}
e_3 - e_2 = e_3(1,\tau) - e_2(1,\tau) \;&=\; \wp(\tfrac{1}{2}+\tfrac{1}{2}\tau) - \wp(\tfrac{1}{2}\tau)\\
&= \sum_{(m,n)=-\infty}^{\infty}\left[\frac{1}{[(m+\frac{1}{2})+(n+\frac{1}{2})\tau]^2} - \frac{1}{[m+(n+\frac{1}{2})\tau]^2}\right];\\
e_1 - e_2 = e_1(1,\tau) - e_2(1,\tau) \;&=\; \wp(\tfrac{1}{2}) - \wp(\tfrac{1}{2}\tau)\\
&= \sum_{(m,n)=-\infty}^{\infty}\left[\frac{1}{[(m+\frac{1}{2})+n\tau]^2} - \frac{1}{[m+(n+\frac{1}{2})\tau]^2}\right].
\end{aligned}$$

The perfect tools for use here are the identities

$$\sum_{m=-\infty}^{\infty}\frac{1}{(m+u)^2} = \frac{\pi^2}{(\sin\pi u)^2}; \tag{17.3.1}$$

$$\sum_{m=-\infty}^{\infty}\frac{1}{(m+\frac{1}{2}+u)^2} = \frac{\pi^2}{(\cos\pi u)^2}. \tag{17.3.2}$$

In fact the difference between the left side and the right side of (17.3.1) or (17.3.2) is an entire function that is periodic, with period 1, that can easily be shown to have limit 0 as $u \to \infty$ in the strip $|\operatorname{Re}u| \leq \frac{1}{2}$. Applying (17.3.1) and (17.3.2) to the previous sums, we get

$$e_3(1,\tau) - e_2(1,\tau) = \pi^2 \sum_{n=-\infty}^{\infty} \left[ \frac{1}{\cos^2 \pi(n+\frac{1}{2})\tau} - \frac{1}{\sin^2 \pi(n+\frac{1}{2})\tau} \right]$$

$$= 2\pi^2 \sum_{n=0}^{\infty} \left[ \frac{1}{\cos^2 \pi(n+\frac{1}{2})\tau} - \frac{1}{\sin^2 \pi(n+\frac{1}{2})\tau} \right]; \quad (17.3.3)$$

$$e_1(1,\tau) - e_2(1,\tau) = \pi^2 \sum_{n=-\infty}^{\infty} \left[ \frac{1}{\cos^2 \pi n\tau} - \frac{1}{\sin^2 \pi(n+\frac{1}{2})\tau} \right]. \quad (17.3.4)$$

Now for real $M$, as $\operatorname{Im}\tau \to \infty$

$$|\cos(M\tau)|^2 = O\left(e^{2|M|\operatorname{Im}\tau}\right); \qquad |\sin(M\tau)|^2 = O\left(e^{2|M|\operatorname{Im}\tau}\right).$$

Therefore the summand indexed by $n$ in the first series in (17.3.4) is dominated by $\exp(-2\pi|n|\operatorname{Im}\tau)$ for large $\operatorname{Im}\tau$. The summands indexed by $n$ in the remaining series in (17.3.3) and (17.3.4) are dominated by $\exp(-\pi|2n+1|\operatorname{Im}\tau)$. It follows from an examination of the associated series that as $\operatorname{Im}\tau \to \infty$,

$$e_1(1,\tau) - e_2(1,\tau) = \pi^2 + O(e^{-\pi\operatorname{Im}\tau}),$$
$$e_3(1,\tau) - e_2(1,\tau) = O(e^{-\pi\operatorname{Im}\tau}).$$

Therefore

$$\lim_{\operatorname{Im}\tau\to\infty} \lambda(\tau) = 0. \quad (17.3.5)$$

We use (17.3.5) with (17.2.4) and (17.2.5) to conclude that

$$\lim_{\tau\to 0} \lambda(\tau) = 1; \qquad \lim_{\tau\to 1} \lambda(\tau) = \infty. \quad (17.3.6)$$

We need a more precise form of (17.3.5), obtained by taking into account the principal terms in (17.3.3), those with $n = 0$ and $n = -1$. These sum to

$$\frac{2\pi^2}{\cos^2(\pi\tau/2)} - \frac{2\pi^2}{\sin^2(\pi\tau/2)} = 2\pi^2 \left\{ \frac{4e^{i\pi\tau}}{(1+e^{i\pi\tau})^2} + \frac{4e^{i\pi\tau}}{(1-e^{i\pi\tau})^2} \right\}.$$

Therefore

$$\lambda(\tau) = 16\, e^{i\pi\tau} + O(e^{-2\pi\operatorname{Im}\tau}) \quad \text{as } \operatorname{Im}\tau \to +\infty. \quad (17.3.7)$$

Note that $\operatorname{Re}\tau = 0$ implies that $e^{i\pi\tau}$ is real and positive, so $\lambda$ is real and positive on the positive imaginary axis.

To complete our discussion of $\lambda$, we consider its behavior on the closure of the domain $\Omega$,

$$\Omega = \{\tau : \operatorname{Im}\tau > 0,\ 0 < \operatorname{Re}\tau < 1,\ |\tau - \tfrac{1}{2}| > \tfrac{1}{2}\}; \quad (17.3.8)$$

see Figure 17.2. The boundary $\Gamma$ of $\Omega$ consists of the three pieces

$$\Gamma_1 = \{\tau : \tau = it,\ 0 < t < \infty\},$$
$$\Gamma_2 = \{\tau : |\tau - \tfrac{1}{2}| = \tfrac{1}{2},\ \operatorname{Im}\tau > 0\},$$
$$\Gamma_3 = \{\tau : \tau = 1 + it,\ 0 < t < \infty\}.$$

Orient the boundary so that $\Omega$ lies to the left.

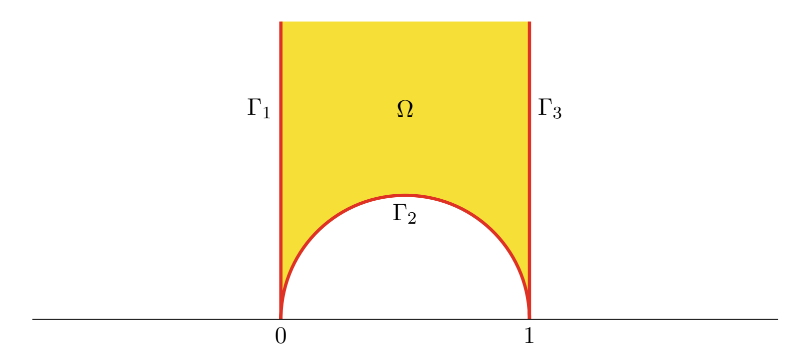



**Fig. 17.2** The domain $\Omega$

The following facts are left as Exercises 3 and 4.

$$\tau \to \frac{1}{1-\tau} \Rightarrow \Omega \to \Omega, \quad \Gamma_1 \to \Gamma_2, \quad \Gamma_2 \to \Gamma_3;$$
$$\tau \to 1 - \frac{1}{\tau} \Rightarrow \Omega \to \Omega, \quad \Gamma_1 \to \Gamma_3, \quad \Gamma_2 \to \Gamma_1; \tag{17.3.9}$$

and

$$\tau \in \Gamma_1 \Rightarrow 0 < \lambda(\tau) < 1, \quad \lim_{\tau\to 0} \lambda(\tau) = 1, \quad \lim_{\tau\to\infty} \lambda(\tau) = 0;$$
$$\tau \in \Gamma_2 \Rightarrow 1 < \lambda(\tau) < \infty, \quad \lim_{\tau\to 0} \lambda(\tau) = 1, \quad \lim_{\tau\to 1} \lambda(\tau) = \infty;$$
$$\tau \in \Gamma_3 \Rightarrow \lambda(\tau) < 0, \quad \lim_{\tau\to 1} \lambda(\tau) = -\infty, \quad \lim_{\tau\to\infty} \lambda(\tau) = 0. \tag{17.3.10}$$

**Theorem 17.3.1.** *The elliptic modular function $\lambda$ is a bijective conformal map of $\Omega$ onto the upper half plane $\mathbb{C}_+$. It extends continuously to the boundary, and $\lambda(\infty)=0$, $\lambda(0) = 1$, $\lambda(1) = \infty$.*

*Proof:* By the above remarks, $\lambda$ extends to the points $0, 1, \infty$. By (17.3.10), $\lambda$ extends across the $\Gamma_j$. Now suppose $w \in \mathbb{C}_+$. Choose a large value $r > \operatorname{Im} w$, and let $\Omega_r$ denote the truncation of the domain $\Omega$ by cutting off the part above the segment $I_r = \{\sigma + ir, 0 < \sigma < 1\}$ and the parts below the image of $I_r$ under the maps $\tau \to 1/(1-\tau)$ and $\tau \to 1 - 1/\tau$. These images are arcs of the circles of radius $1/2r$ with centers $i/2r$ and $1 + i/2r$, respectively. As $\tau$ follows the oriented boundary of $\Omega_r$ starting with the point $ir$, $\lambda$ starts from a positive value that is close to zero, passes through real values to values close to 1, then moves through positive values along part of $\Gamma_2$. Along the image of $I_r$ near $\tau = 1$ we have values $\lambda(1 - 1/\tau)$, $\tau \in I_r$.

As noted in (17.2.8), $\lambda(1-1/\tau) = 1-1/\lambda(\tau)$. In view of this and (17.3.7), as $\tau$ runs along $I_r$, $\lambda(\tau)$ is close to the large semicircle that runs from $e^{\pi r}/16$ to $-e^{\pi r}/16$ through $ie^{\pi r}/16$. Continuing up the relevant portion of $\Gamma_3$, $\lambda$ reaches a small negative value. On $I_r$, the image is close to a semicircle of radius $1/2r$, ending with a small positive value. Thus $\lambda$ maps the interior of the cutoff domain $\Omega_r$ onto a domain that, as $r \to \infty$, covers $\mathbb{C}_+$, and the map is bijective. □

**Theorem 17.3.2.** *The elliptic modular function $\lambda$ is a countably-many to one conformal map of $\mathbb{C}_+$ onto $\mathbb{C} \setminus \{0,1\}$.*

*Proof:* We know that $\lambda$ is a holomorphic map from $\mathbb{C}_+$ to $\mathbb{C} \setminus \{0,1\}$, and that $\lambda$ is conformal on $\Omega$, continous on the closure of $\Omega$, and real on the boundary of $\Omega$. Therefore, on the image of $\Omega$ under the reflection across a piece $\Gamma_j$ of the boundary, $\lambda$ is given by the reflection of $\lambda$ on $\Omega$ and is a conformal map onto $\mathbb{C}_-$. This boundary-crossing procedure can be continued across each of the smooth parts of the boundary of the image of $\Omega$, leading to conformal maps to $\mathbb{C}_+$, and so on. Because we already have $\lambda$ as a globally defined function on $\mathbb{C}_+$, we know that all these continuations are simply part of $\lambda$. The only fact that needs to be checked is that the union of all these reflections, together with their boundaries in $\mathbb{C}_+$, is all of $\mathbb{C}_+$. Since reflections across vertical boundaries mean that the union under consideration is invariant under translation by 1, the question is what happens under reflections through the curved boundaries.

Reflection through a circle of radius $r$ with center $p \in \mathbb{R}$ is given by the map

$$z \;\to\; p + \frac{r^2}{\bar{z} - p} \tag{17.3.11}$$

which fixes the circle $|z-p| = r$ and maps $\infty$ to $p$. This map is a linear fractional transformation followed by complex conjugation, so it takes the vertical lines $\{z : \mathrm{Re}\, z = p \pm r\}$ to the circles $|z-(p \pm \frac{1}{2}r)| = \frac{1}{2}r$. The process starts with reflection across a lower boundary made up of semicircles with radius $1/2$ and centers at $p = n + \frac{1}{2}$, $n$ any integer. After $k$ iterations, the lower boundary is formed by non-overlapping semicircles with radius $2^{-k-1}$ and centers on the real line. Therefore the union of all the images is all of $\mathbb{C}_+$. □

**Remark.** In view of Theorems 17.3.1 and 17.3.2, the results of Section 6.5 imply that $\lambda$ is a quotient of two solutions of the hypergeometric equation (6.5.4). See Exercise 18 of Chapter 6.

The exponential function is an example of an entire function that takes all values except 0. (In fact each such function is a constant times the exponential of an entire function.)

**Theorem 17.3.3.** (Picard) *If an entire function $f$ omits two values, it is constant.*

*Proof:* Suppose that $f(z)$ is never $a$ or $b$, with $a \neq b$. Let

$$g(z) = \frac{f(z)-a}{f(z)-b}.$$

Then $g$ omits 0 and 1. It follows that the composition $\lambda^{-1} \circ g$ is well-defined, though multi-valued. However we may choose a single-valued branch. If $D_r(0)$ is a sufficiently small disk centered at the origin, we may choose a branch $G_0$ of $\lambda^{-1} \circ g$ that is defined and holomorphic in the disk. The function $G_0$ can be continued analytically along any curve in $\mathbb{C}$. Since $\mathbb{C}$ is simply connected, the monodromy theorem, Theorem 1.7.2, says that these continuations define an entire function $G$. Since the range of $G$ is included in $\mathbb{C}_+$, it follows that $\exp(iG)$ is bounded, hence constant. Therefore $G$ is constant, so $g = \lambda \circ G$ is constant, so $f$ is constant. □

**Remarks.** Theorem 17.3.3 is often called Picard's "little" theorem, in contrast to Picard's "big" theorem, which he proved later: a function with an essential singularity cannot omit more than one point. This requires a somewhat closer look at the process of successive reflection across edges. The proof is outlined in Exercises 12–24.

## 17.4 Automorphic functions; the *J* function

An *automorphic function* of one complex variable is a meromorphic function $f$, defined on a domain $\Omega$, that is invariant under some infinite discrete group of linear fractional transformations that map $\Omega$ onto itself:

$$f(g(z)) = f(z), \qquad z \in \Omega, \quad g \in G.$$

Examples, with $\Omega = \mathbb{C}$, are trigonometric functions and elliptic functions, invariant under one or two groups of translations.

The elliptic modular function $\lambda$, defined on $\mathbb{C}_+$, has period 1 and transforms nicely under the modular group. Moreover, $\lambda$ is invariant under those transformations in the modular group that are equivalent mod 2 to the identity transformation, such as

$$g(\tau) = \frac{5\tau+4}{6\tau+5},$$

so $\lambda$ is a different type of automorphic function, invariant under a non-commutative infinite transformation group.

To find a function invariant under the *full* modular group, we look for a function that, like $\lambda$, is a function of the roots $e_j$. For full invariance, we need a symmetric function of the roots. The simplest such functions are the elementary symmetric polynomials

$$e_1 + e_2 + e_3 = 0, \qquad 4(e_1e_2 + e_2e_3 + e_3e_1) = -g_2, \qquad 4e_1e_2e_3 = g_3.$$

If we want a function that is a function of $\tau = \omega_2/\omega_1$, we can do as before and take advantage of the homogeneity property of the $e_j$ by arranging a quotient of expressions that have the same homogeneity. If we try this with $g_2$, which has degree $-4$, and $g_3$, which has degree $-6$, we are led to consider $g_3^2$ and $g_2^3$, each of degree $-12$. Thus it can be argued that the simplest such functions of $\tau$ have the form

$$F(\tau) \;=\; \frac{ag_2^3 + bg_3^2}{cg_2^3 + dg_3^2}, \qquad \operatorname{Im}\tau > 0. \tag{17.4.1}$$

This function will be holomorphic in $\mathbb{C}_+$ if the denominator is never zero. Now the fact that the $e_j$ are distinct means that the discriminant of the polynomial $Q(t) = 4t^3 - g_2t - g_3$ does not vanish. This leads us to take as denominator the discriminant $g_2^3 - 27g_3^2$, and a numerator as simple as possible. The result is the $J$-function

$$J(\tau) \;=\; \frac{g_2^3}{g_2^3 - 27g_3^2}. \tag{17.4.2}$$

Arguments similar to the one used for $\lambda$ show that the set $\Delta$ of (17.2.2) is a fundamental domain for $J$, and that $J$ is a conformal map from the closure of $\Delta$ onto $\mathbb{C}_+$.

Another way to create a function that is symmetric in the roots is to take all the forms of $\lambda$ that are obtained by the action of the modular group and take a symmetric function of these forms. They can be enumerated as coming from

$$\frac{e_3 - e_2}{e_1 - e_2}, \quad \frac{e_1 - e_3}{e_2 - e_3}, \quad \frac{e_2 - e_1}{e_3 - e_1}, \quad \frac{e_1 - e_2}{e_3 - e_2}, \quad \frac{e_2 - e_3}{e_1 - e_3}, \quad \frac{e_3 - e_1}{e_2 - e_1}.$$

In terms of $\lambda$ these are, respectively,

$$\lambda, \quad \frac{\lambda - 1}{\lambda}, \quad \frac{1}{1 - \lambda}, \quad \frac{1}{\lambda}, \quad \frac{\lambda}{\lambda - 1}, \quad 1 - \lambda.$$

One of the simplest symmetric functions is obtained by adding 1 to each term and taking the product:

$$\begin{aligned} F(\tau) &= (1+\lambda)\left(1 + \frac{\lambda - 1}{\lambda}\right)\left(1 + \frac{1}{1-\lambda}\right)\left(1 + \frac{1}{\lambda}\right)\left(1 + \frac{\lambda}{\lambda - 1}\right)(1 + 1 - \lambda) \\ &= -\frac{[(-3e_2)(-3e_3)(-3e_1)]^2}{[(e_1 - e_2)(e_2 - e_3)(e_3 - e_1)]^2} \\ &= = -\frac{3^6 4^{-2} g_3^2}{[(e_1 - e_2)(e_2 - e_3)(e_3 - e_1)]^2}. \end{aligned} \tag{17.4.3}$$

After some calculation, using $e_1 + e_2 + e_3 = 0$, the denominator is

$$-27(e_1e_2e_3)^2 - 4(e_1e_2 + e_2e_3 + e_3e_1)^3 \;=\; \frac{g_2^3 - 27g_3^2}{16}. \tag{17.4.4}$$

Thus

$$F(\tau) \;=\; \frac{(27)^2 g_3^2}{27g_3^2 - g_2^3} \;=\; 27[1 - J(\tau)], \tag{17.4.5}$$

so

$$J(\tau) \;=\; 1 - \frac{F(\tau)}{27} \;=\; \frac{4}{27}\,\frac{(1-\lambda+\lambda^2)^3}{\lambda^2(1-\lambda)^2}. \tag{17.4.6}$$

Since the function $J(\tau)$ has period 1, we expect it to have a Fourier expansion

$$J(\tau) \;=\; \sum_{-\infty}^{\infty} a_n t^n, \qquad t = e^{2\pi i\tau}.$$

When $\tau$ is in $\mathbb{C}_+$, $|t| < 1$. Therefore we expect only finitely many negative powers of $t$. In fact it is known that

$$J(\tau) \;=\; \frac{1}{1728}\left[\frac{1}{t} + 744 + 196884t + 21493760t^2 + \dots\right]. \tag{17.4.7}$$

Here we sketch the calculation of the first two terms of this expansion. The starting point for calculating (17.4.7) is the calculation of the Fourier expansions of

$$g_2(\tau) \;=\; 60 \sum_{(m,n)\neq(0,0)} \frac{1}{(m+n\tau)^4}, \qquad g_3 \;=\; 140 \sum_{(m,n)\neq(0,0)} \frac{1}{(m+n\tau)^6}. \tag{17.4.8}$$

Now

$$\begin{aligned}\sum_{(m,n)\neq(0,0)}^{\infty} \frac{1}{(m+n\tau)^{2k}} &= 2\sum_{m=1}^{\infty}\frac{1}{m^{2k}} + \sum_{m=-\infty}^{\infty}\sum_{n\neq 0}\frac{1}{(m+n\tau)^{2k}} \\ &= 2\zeta(2k) + 2\sum_{n=1}^{\infty}\sum_{m=-\infty}^{\infty}\frac{1}{(m+n\tau)^{2k}}.\end{aligned} \tag{17.4.9}$$

The values of the (Riemann) $\zeta$-function here are

$$\zeta(4) \;=\; \frac{\pi^4}{90}, \qquad \zeta(6) \;=\; \frac{\pi^6}{945}. \tag{17.4.10}$$

Calculation of the corresponding sums in (17.4.8) starts with the identity

$$\begin{aligned}\pi\cot\pi u &= \frac{1}{u} + \sum_{n=1}^{\infty}\left(\frac{1}{u-n} + \frac{1}{u+n}\right) \\ &= \lim_{N\to\infty}\sum_{n=-N}^{N}\frac{1}{u-n}.\end{aligned} \tag{17.4.11}$$

In fact the difference between the two sides is entire, periodic with period 1, and vanishes as $u \to \infty$ in a period strip like $\{z : |\mathrm{Re}\, z| \le 1/2\}$, so the difference is zero. Now $u \in \mathbb{C}_+$ implies that $|e^{i\pi u}| < 1$, so

$$\begin{aligned}\pi \cot \pi u &= \pi i \, \frac{e^{i\pi u} + e^{-i\pi u}}{e^{i\pi u} - e^{-i\pi u}} \\ &= \pi i \, \frac{w+1}{w-1} \\ &= -\pi i \, (1+w)(1+w+w^2+\dots) \\ &= -\pi i \, (1+2w+2w^2+\dots), \quad w = e^{2i\pi u}. \end{aligned} \tag{17.4.12}$$

Combining (17.4.11) and (17.4.12) and differentiating gives

$$6 \sum_{m=-\infty}^{\infty} \frac{1}{(m+u)^4} = 16\pi^4 (w + 8w^2 + 27w^3 + \dots); \tag{17.4.13}$$

$$120 \sum_{m=-\infty}^{\infty} \frac{1}{(m+u)^6} = -64\pi^6 (w + 32w^2 + 243w^3 + \dots). \tag{17.4.14}$$

We can now combine (17.4.8), (17.4.9), (17.4.10), (17.4.13), and (17.4.14) to obtain

$$g_2(\tau) = \frac{4\pi^4}{3} [1 + 240\,(t + 9t^2 + \dots)]; \tag{17.4.15}$$

$$g_3(\tau) = \frac{8\pi^6}{27} [1 - 504\,(t + 33t^2 + \dots)]. \tag{17.4.16}$$

In turn, these lead to

$$J(\tau) = \frac{1 + 720t + \dots}{1728(t - 24t^2 + \dots}), \tag{17.4.17}$$

which gives the first two terms of the expansion (17.4.7).

## Exercises

1. Check the various assertions in the proof of Proposition 17.2.1.
2. Show that $R$ and $T$ generate the modular group. (Hint: use $R$, $T$, and $T^{-1}$ to reduce the largest absolute value among the entries of a given $g$.)
3. Verify (17.3.9). (Since lines or circles are taken to lines or circles, it is enough to track what happens to the points $0$, $\infty$, $1+i$, $1$, $(1+i)/2$, and $i$.)
4. Verify the assertions in (17.3.10). (Note that $\lambda$ is positive on the positive imaginary axis and is never 0 or 1.)
5. Show that if $F$ has the form (17.4.1) and has no poles in $\mathbb{C}_+$, $F$ must be of the form $F(\tau) = a_1 J(\tau) + b_1$.
6. Verify (17.4.4).

7. With $P(z) = 4z^3 - g_2 z - g_3$, find polynomials $Q_j$, $P_j$ such that the $Q_j$ have positive leading coefficient and

$$P(z) \;=\; P'(z)Q_1(z) + P_1(z); \qquad P'(z) \;=\; P_1(z)Q_1(z) + P_0(z),$$

where $P_j$ has degree $j$. Show that a polynomial that divides both $P$ and $P'$ also divides $P_1$ and $P_0$. Conclude that $P$ has a multiple root if and only if $P_0 = 0$. This gives a second derivation of the discriminant.

8. Verify (17.4.13) and (17.4.14).
9. Verify (17.4.15) and (17.4.16).
10. Verify (17.4.17).
11. Verify:

$$21493760 = 1 + 196883 + 21296876,$$
$$864299970 = 2\cdot 1 + 2\cdot 196883 + 21296876 + 842609326.$$

(See (17.4.7) and the Addendum.)

The remaining exercises outline Picard's proof of his "big" theorem: a function with an essential singularity cannot omit more than one finite value.

12. Use Exercise 2 and (17.2.4), (17.2.5) to show that if $B$ belongs to the modular group, then

$$\lambda \circ B(z) \;=\; \frac{a\lambda(z) + b}{c\lambda(z) + d}$$

where $a, b, c, d$ are integers and $ad - bc = 1$ if $\operatorname{Im}\lambda(z)$ and $\operatorname{Im}\lambda(B(z))$ have the same sign, and $ad - bc = -1$ otherwise.

13. Show that each of the reflections that are used to extend $\Omega$ so as to cover all of $\mathbb{C}_+$ has the form

$$z \;\to\; \frac{a\bar{z} + b}{c\bar{z} + d}, \tag{17.4.18}$$

where $a, b, c, d$ are integers and $ad - bc = -1$. (In fact the reflections through circular arcs, given by (17.3.11), start with $r = 1/2$ and $2p$ an integer, and subsequent reflections have $r = 2^{-k}$ and $2^k p$ an integer. Also, reflection through a line $\{z : \operatorname{Re} z = n\}$ is given by

$$z \;\to\; 2n - \bar{z}.)$$

14. Show that the composition of two of the reflections from Exercise 13 belongs to the modular group.
15. Suppose $\Omega_1$ is one of the domains that is obtained from $\Omega$ by an even number of reflections through sides. Show that the restrictions of $\lambda$ are related by

$$\lambda\big|_{\Omega_1} \;=\; A\left(\lambda\big|_{\Omega}\right) \tag{17.4.19}$$

where $A$ is a member of the modular group.

16. Suppose that $\Omega_1$ and $\Omega_2$ are domains obtained from $\Omega$ by an odd number of reflections through sides. Show that the respective restrictions of $\lambda$ are related by

$$\lambda|_{\Omega_2} = A\left(\lambda|_{\Omega_1}\right) \tag{17.4.20}$$

where $A$ is a member of the modular group.

17. Recall that we are considering the modular group to consist either of matrices or of the corresponding linear fractional transformations; the context should make clear which is meant. Suppose that the matrix $A$ belongs to the modular group and is not the identity or its negative. Show that the sum of the eigenvalues is real, and there are exactly three possibilities:

   (a) $A$ has two complex eigenvalues $e^{i\theta}$, $e^{-i\theta}$, $0 < \theta < \pi$.

   (b) $A$ has two distinct real eigenvalues $\mu$, $1/\mu$.

   (c) $A$ has a single eigenvalue 1 or $-1$.

18. With $A$ as in the Exercise 17, show that in case (a) or case (b), there is a matrix $S$ such that $S^{-1}AS$ is diagonal. Show that in case (c) there is a matrix $S$ such that $S^{-1}AS = T$, where $T$ is given in (17.2.3). In case (a), one may assume $S(\mathbb{C}_+) = \mathbb{D}$, the unit disk; in cases (b) and (c), one may take $S(\mathbb{C}_+) = \mathbb{C}_+$. Note that $S^{-1}AS = B$ is equivalent to $AS = SB, \det S \neq 0$.

19. Suppose that $f$ has an essential singularity at a point $z_0$. Suppose that $f$ does not take two distinct values $a$, $b$. Show that there is a function $g$ with an essential singularity at $\infty$ such that $g$ does not take the values 0, 1.

20. Suppose that $g$ is holomorphic in a neighborhood $N = \{z : |z| > R\}$ of $\infty$ and does not take the values 0 or 1. Choose a determination $G$ of $\lambda^{-1} \circ g$ that is holomorphic in the (simply connected) domain $N_1 = \{z : |z| > R, z \notin [0,\infty)\}$.

   (a) Suppose $z \in N_1$. Let $G_+(z)$ be the value obtained by continuing $G$ in the positive direction from $z$ around the circle of radius $|z|$ until one returns to $z$. Show that

$$G_+(z) = A\,G(z), \quad \text{for some } A \text{ in the modular group.}$$

   (b) Show that $A$ in (a) is the same for every $z \in N_1$. (In fact $A$ depends continuously on $z$, but elements of the modular group are isolated points.)

21. Suppose that $A$ in Exercise 20 is the identity. Show $\infty$ is not an essential singularity of $g$, e.g. via the Casorati–Weierstrass theorem.

22. Suppose that $A$ in Exercise 20 is of type (a) from Exercise 17, and choose $S$ so that $S^{-1}AS$ is diagonal. This can be done so that

$$S\,G_+ = e^{2i\theta} S\,G, \quad 0 < \theta < 2\pi.$$

   Let $H(z) = z^{-\theta/\pi} S\,G(z)$. Show that $H$ is bounded and holomorphic in $N = \{z : |z| > R\}$ and vanishes at $\infty$. Use this to conclude that $\infty$ is not an essential singularity of $g = \lambda \circ G$.

23. Suppose that $A$ in Exercise 20 is of type (b) from Exercise 17, and choose $S : \mathbb{C}_+ \to \mathbb{D}$ so that $S^{-1}AS$ is diagonal. This can be done so that

$$SG_+ = \mu^2 SG, \quad \mu > 1.$$

Let $G_{[n]}$ denote the result of continuing $G$ along $n$ counterclockwise circuits around the origin: $G_{[2]} = (G_+)_+$, etc. Show that

$$\mu^{2n} SG = [SG]_{[n]} = z^{n \log \mu / \pi i} H(z),$$

where $H$ is holomorphic in $N$. Choose $m$ large enough so that $z_0 = e^{m/\log \mu}$ belongs to $N$. Then

$$\mu^{2n} SG(z_0) = e^{mn/\pi i} H(z_0).$$

The left side is in $\mathbb{C}_+$; show that the right side cannot be in $\mathbb{C}_+$ for any $n$.

24. Suppose, finally, that $A$ in Exercise 20 is of type (c) from Exercise 17. Then there is $S : \mathbb{C}_+ \to \mathbb{C}_+$ such that

$$SG_+ = SG + 1.$$

Show that $H(z) = SG(z) - \log z / 2\pi i$ is holomorphic in $N$. Then

$$|\exp(iH(z))| = |z^{-1/2\pi} \exp(iSG(z))| \le |z|^{-1/2\pi},$$

so $\operatorname{Im} H(z) \ge 0$ in $N$, which implies that $\infty$ is not an essential singularity of $H$ (Casorati–Weierstrass). Use this to conclude that $\infty$ cannot be an essential singularity of $g = \lambda \circ G$. This is Picard's "big" theorem.

## Addendum: Moonshine

There is a renormalized version of $J$:

$$j(\tau) = (12)^3 J(\tau) - 744 = \frac{(12)^3 g_2(\tau)^3}{g_2(\tau)^3 - 27 g_3(\tau)^2} - 744.$$

Since $j(\tau)$ has period 1 and is bounded as $\operatorname{Im} \tau \to +\infty$, it has a Fourier expansion

$$j(\tau) = q^{-1} + 196884\, q + 21493760\, q^2 + 864299970 q^3 + 20245856256\, q^4 + \dots,$$

where $q = q(\tau) = e^{2\pi i \tau}$. (The reason for multiplying by $(12)^3$ in the first equation is to make the leading term $q^{-1}$ rather than $q^{-1}/1728$. The reason for subtracting 744 is to kill the constant term in the expansion; obviously these modifications do not affect the group invariance property). The functions $J$ and $j$ have long been known to number theorists and algebraic geometers.

Group theorists spent decades, starting in the 1960s, classifying the finite simple groups. These fall into several infinite families, together with 26 others – the *spo-*

*radic groups*. The largesty of the sporadic groups, known as the "monster," has order

$$2^{46} \cdot 3^{20} \cdot 5^{9} \cdot 7^{6} \cdot 11^{2} \cdot 13^{3} \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71.$$

It has irreducible representations on spaces of dimension

$$1, \quad 196883, \quad 21296876, \quad 842609326, \quad 18538750076, \quad 19360062527, \quad \dots .$$

It was noted by McKay that

$$196884 = 1 + 196883,$$

and then by Thompson that

$$\begin{aligned}
21493760 &= 1 + 196883 + 21296876, \\
864299970 &= 2 \cdot 1 + 2 \cdot 196883 + 21296876 + 842609326, \\
20245856256 &= 2 \cdot 1 + 3 \cdot 196883 + 2 \cdot 21296876 \\
&\quad + 842609326 + 19360062527.
\end{aligned}$$

There was a general feeling among group theorists and others that these are not simply numerical coincidences. For (much) more on this and related topics, such as vertex algebras and quantum gravity, see Duncan, Griffin, and Ono [39].

## Remarks and further reading

The theory of automorphic functions and automorphic forms is a large and active area of current research, with far-reaching applications, among them the Langlands program: see the references at the end of Chapter 12.

For classical connections to complex analysis, see Siegel [128]. For much further study of the theory in one complex variable, see Lehner [85], [86]. For other modern connections, see Dou and Zhang [37], Lozano-Robledo [94] and Venkov [137]. For some history, see Roy [122].

# Chapter 18
# Integral transforms




The Cauchy integral formula

$$f(z) = \frac{1}{2\pi i} \int_{\Gamma} \frac{f(\zeta)}{\zeta - z} d\zeta,$$

gives the value at $z$ of a function $f$ that is holomorphic in a domain that contains $z$ and is continuous on the boundary $\Gamma$. This formula can be considered as a particular case of an integral transform that takes a given function $f$ defined on a complex curve $\Gamma$ to the function

$$C_{\Gamma} f(z) = \frac{1}{2\pi i} \int_{\Gamma} \frac{f(\zeta)}{\zeta - z} d\zeta, \quad z \notin \Gamma. \tag{18.0.1}$$

What happens if $f$ is *not* the boundary value of a function that is holomorphic inside the curve? This question leads naturally to the study of a second integral transform associated with the curve.

In the first section we introduce the concept of an approximate identity and also a special class of functions that are well-adapted to the consideration of various integral transforms.

In the second section we consider two integral transforms. One is the Cauchy transform (18.0.1) when $\Gamma = \mathbb{R}$. The second is the Hilbert transform, which is related to the values of the Cauchy transform at $\mathbb{R}$.

These considerations lead naturally to the Fourier transform, and to the spaces $L^1$ and $L^2$, which are the subject of the remaining sections.

## 18.1 Approximate identities and Schwartz functions

A useful tool for approximation of functions is an *approximate identity*. This occurs implicitly in Section 4.2 for the case of the interval $(-\pi, \pi)$. In the case of the line, an approximate identity is a family of continuous functions $\{G_\varepsilon\}_{0<\varepsilon\leq 1}$, or a



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_18

sequence of such functions $\{G_n\}_{n=1}^{\infty}$, with two properties. We state the properties here for the case $\varepsilon \to 0$; the reader may easily translate to the case $n \to \infty$.

$$G_\varepsilon(x) \geq 0; \tag{18.1.1}$$

$$\int_{-\infty}^{\infty} G_\varepsilon(x)\,dx = 1, \tag{18.1.2}$$

and for each $\delta > 0$,

$$\lim_{\varepsilon \to 0} \int_{|x|>\delta} G_\varepsilon(x)\,dx \;=\; 0. \tag{18.1.3}$$

Thus, as $\varepsilon \to 0$, $G_\varepsilon$ becomes more and more concentrated near the origin.

The example

$$G_\varepsilon(x) \;=\; \frac{\varepsilon}{\pi(x^2+\varepsilon^2)} \tag{18.1.4}$$

will occur in connection with the Cauchy transform. A second, much-used example is the family of Gaussian probability densities

$$G_\varepsilon(x) \;=\; \frac{e^{-x^2/2\varepsilon}}{\sqrt{2\pi\varepsilon}}. \tag{18.1.5}$$

Many function spaces can be described as the completion of some class of functions $\mathcal{F}$ with respect to a metric

$$d(f,g) \;=\; ||f-g||$$

defined by a norm $||\cdot||$. In this chapter we consider the analogue for $\mathbb{R}$ of the functions on the interval $(-\pi,\pi)$ that are considered in Section 1.10. The norms in this case are

$$||f||_p \;=\; \left[\int_{-\infty}^{\infty} |f(x)|^p\,dx\right]^{1/p}, \qquad 1 \leq p < \infty.$$

We take the class $\mathcal{F}$ to consist of continuous $f : \mathbb{R} \to \mathbb{C}$ such that $f$ vanishes outside some bounded interval (depending on $f$). For $1 \leq p < \infty$ the completion is the $L^p$ space $L^p(\mathbb{R})$. Similarly the space $C_0(\mathbb{R})$ of continuous functions $f$ with limit 0 as $|x| \to \infty$ is the completion of $\mathcal{F}$ with respect to the norm

$$||f||_\infty \;=\; \sup_{x\in\mathbb{R}} |f(x)|.$$

Suppose that $f$ and $g$ are two continuous functions from $\mathbb{R}$ to $\mathbb{C}$, at least one of which vanishes outside some bounded interval. The *convolution* $f * g$ is the function defined by

$$f*g(x) \;=\; \int_{-\infty}^{\infty} f(x-y)\,g(y)\,dy. \tag{18.1.6}$$

Convolution is commutative:

$$f*g(x) \;=\; g*f(x); \tag{18.1.7}$$

see Exercise 1.

**Proposition 18.1.1.** *Suppose that* $\{G_\varepsilon\}$ *is an approximate identity and that* $f : \mathbb{R} \to \mathbb{C}$ *is continuous and has limits* $f(x) \to a_\pm$ *as* $x \to \pm\infty$*. Then the convolution* $f * G_\varepsilon$ *converges uniformly to* $f$ *as* $\varepsilon \to 0$*.*

*Proof:* Properties (18.1.1) and (18.1.2), imply that

$$|f * G_\varepsilon(x) - f(x)| = \left| \int_{-\infty}^{\infty} [f(x-y) - f(x)] G_\varepsilon(y)\, dy \right|$$
$$\leq \sup_{|y| \leq \delta} |f(x-y) - f(x)| + 2 \sup_x |f(x)| \int_{|y|>\delta} G_\varepsilon(y)\, dy.$$

Under our assumption, $f$ is bounded and uniformly continuous. Therefore the first term in the last line is small for small $\delta$, uniformly with respect to $x$, while property (18.1.3) implies that for a given $\delta$, the second term is small for small $\varepsilon$, uniformly with respect to $x$. □

Functions of the *Schwartz class* $\mathcal{S}$ are particularly useful in connection with the study of the Fourier transform. A function $f : \mathbb{R} \to \mathbb{C}$ belongs to $\mathcal{S}$ if each derivative is continuous and decays faster than each power of $1/|x|$ as $|x| \to \infty$. This means that for each $k$ and $n$,

$$\sup_{x \in \mathbb{R}} |x|^n |f^{(k)}(x)| < \infty. \tag{18.1.8}$$

The Gaussian functions (18.1.5) are examples. In fact

$$e^{x^2/2\varepsilon} > \frac{x^{2n}}{(2\varepsilon)^n\, n!}$$

so

$$e^{-x^2/2\varepsilon} = O(x^{-2n}) \text{ as } |x| \to \infty, \text{ all } n.$$

Each derivative of $e^{-x^2/2\varepsilon}$ has the form $p(x)e^{-x^2/2\varepsilon}$, where $p$ is a polynomial, so derivatives also satisfy such estimates.

We can use the approximate identity (18.1.5) to approximate by functions in $\mathcal{S}$.

**Theorem 18.1.2.** *Suppose that* $f : \mathbb{R} \to \mathbb{C}$ *is continuous, and vanishes outside a bounded interval. Then there are Schwartz class functions* $f_\varepsilon$ *such that* $f_\varepsilon$ *converges uniformly to* $f$*, as* $\varepsilon \to 0$*, and also converges to* $f$ *with respect to each of the* $L^p$ *norms* $||\ ||_p$ *and the uniform norm* $||\ ||_\infty$*.*

*Proof:* Let $\{G_\varepsilon\}$ be the approximate identity (18.1.5). By Proposition 18.1.1 the functions $f_\varepsilon = G_\varepsilon * f$ converge uniformly to $f$. Suppose that $f(x) = 0$ for $|x| \geq A$. Then

$$f_\varepsilon(A+t) = \int_{-A}^{A} f(y)\, G_\varepsilon(t+A-y)\, dy = \int_{-2A}^{0} f(u+A) G_\varepsilon(t-u)\, du.$$

For $t \geq 0$ and $u \leq 0$, $G_\varepsilon(t-u)$ is maximal at $u = 0$. Therefore for $t \geq 0$,

$$|f_\varepsilon(A+t)| \leq C G_\varepsilon(t) = \frac{C}{\sqrt{2\varepsilon\pi}} e^{-t^2/\varepsilon}, \qquad C = 2A \sup_x |f(x)|. \tag{18.1.9}$$

The same estimate is valid for $|f_\varepsilon(-A-t)|$, $t \geq 0$. The same argument applies to derivatives $[f * G_\varepsilon]^{(n)} = f * (G_\varepsilon^{(n)})$. It follows that each $f_\varepsilon$ belongs to $\mathcal{S}$. Moreover

$$G_\varepsilon(t) \leq \frac{1}{\sqrt{2\varepsilon\pi}} \cdot \frac{2\varepsilon}{t^2} = \left[\frac{2\varepsilon}{\pi}\right]^{1/2} \cdot \frac{1}{t^2}. \tag{18.1.10}$$

We know that $f_\varepsilon$ converges uniformly to $f$ on the interval $[-A-1, A+1]$. It follows from (18.1.9) and (18.1.10) that for each $1 \leq p < \infty$,

$$\lim_{\varepsilon \to 0} \int_{|x|>A+1} |f_\varepsilon(x)|^p \, dx = 0.$$

Therefore the $f_\varepsilon$ converge to $f$ with respect to the $L^p$ norm. The same is true with respect to the norm $|| \, ||_\infty$. □

By definition, $\mathcal{F}$ is dense in $L^p(\mathbb{R})$ and $C_0(\mathbb{R})$, so Theorem 18.1.2 has the following consequence.

**Corollary 18.1.3.** *Schwartz functions are dense in $C_0(\mathbb{R})$ and in $L^p(\mathbb{R})$, $1 \leq p < \infty$.*

## 18.2 The Cauchy Transform and the Hilbert transform

The *Cauchy transform* $Cf$ of $f \in \mathcal{S}$ is defined to be

$$Cf(z) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{f(x)}{x-z} \, dx, \quad z \notin \mathbb{R}, \tag{18.2.1}$$

It is easily checked that $Cf$ is holomorphic in each half plane $\mathbb{C}_\pm = \{z : \pm \text{Im}\, z > 0\}$.

We will analyze the boundary values of $Cf(z)$ as $z$ approaches $\mathbb{R}$ from $\mathbb{C}_+$ or $\mathbb{C}_-$ by considering the two linear combinations

$$Cf(x+i\varepsilon) - Cf(x-i\varepsilon), \tag{18.2.2}$$

and

$$Cf(x+i\varepsilon) + Cf(x-i\varepsilon) \tag{18.2.3}$$

as $\varepsilon \to 0+$.

**Theorem 18.2.1.** *For $f \in \mathcal{S}$ and $z \in \mathbb{C}_+$,*

$$\lim_{z \to x} [Cf(z) - Cf(\bar{z})] = f(x), \quad x \in \mathbb{R}, \tag{18.2.4}$$

*uniformly on* $\mathbb{R}$.

*Proof:* The first step is to prove uniform convergence along vertical lines, i.e. that the difference (18.2.2) converges uniformly to $f(x)$. Given $\varepsilon > 0$,

$$\begin{aligned} Cf(x+i\varepsilon) - Cf(x-i\varepsilon) &= \frac{1}{2\pi i}\int_{-\infty}^{\infty}\left\{\frac{1}{t-x-i\varepsilon} - \frac{1}{t-x+i\varepsilon}\right\} f(t)\,dt \\ &= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\varepsilon f(t)}{(t-x)^2+\varepsilon^2}\,dt \\ &= f * G_\varepsilon(x), \end{aligned}$$

where

$$G_\varepsilon(x) \;=\; \frac{\varepsilon}{\pi(x^2+\varepsilon^2)}. \tag{18.2.5}$$

Now $\{G_\varepsilon\}$ is an approximate identity (Exercise 2). The argument of Proposition 18.1.1 extends easily to the case of $f \in \mathcal{S}$. Thus the difference (18.2.2) converges uniformly to $f$.

An integration by parts shows that the derivative

$$[Cf]'(z) \;=\; \frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{f(x)}{(x-z)^2}\,dx \;=\; \frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{f'(x)}{x-z}\,dx.$$

Thus the derivative of the difference (18.2.2) converges uniformly to $f'$. In particular, the derivative is uniformly bounded. The uniform convergence of (18.2.2) is an easy consequence (Exercise 3). $\square$

As we shall see, examination of the limit of the sum (18.2.3) leads to the *Hilbert transform* of $f \in \mathcal{S}$, defined as a "principal value" integral

$$Hf(x) \;=\; \text{p.v.}\int_{-\infty}^{\infty}\frac{f(x-y)}{y}\,dy \;\equiv\; \lim_{\delta\to 0+}\int_{|y|>\delta}\frac{f(x-y)}{y}\,dy. \tag{18.2.6}$$

The integrand is integrable at $\infty$, so it is enough to look at the integral for $\delta < |y| < 1$. Because $1/y$ is an odd function, its integral over this range is zero. Therefore

$$\int_{\delta<|y|<1}\frac{f(x-y)}{y}\,dy \;=\; \int_{\delta<|y|<1}\frac{f(x-y)-f(x)}{y}\,dy.$$

As a Schwartz function, $f$ has a bounded derivative, so the second integrand is bounded for $|y| \le 1$, uniformly with respect to $x$. Thus the limit (18.2.6) exists.

**Theorem 18.2.2.** *For* $f \in \mathcal{S}$ *and* $z \in \mathbb{C}_+$,

$$\lim_{z\to x}[Cf(z) + Cf(\bar{z})] \;=\; \frac{i}{\pi}Hf(x), \quad x \in \mathbb{R}, \tag{18.2.7}$$

*uniformly on* $\mathbb{R}$.

*Proof:* Again the first step is to prove uniform convergence along vertical lines, i.e. that (18.2.3) converges uniformly. Given $\varepsilon > 0$,

$$\begin{aligned} Cf(x+i\varepsilon)+Cf(x-i\varepsilon) &= \frac{1}{2\pi i}\int_{-\infty}^{\infty}\left\{\frac{1}{t-x-i\varepsilon}+\frac{1}{t-x+i\varepsilon}\right\}f(t)\,dt \\ &= \frac{1}{\pi i}\int_{-\infty}^{\infty}\frac{t-x}{(t-x)^2+\varepsilon^2}\,f(t)\,dt \\ &= \frac{i}{\pi}\int_{-\infty}^{\infty} f(x-y)\,\frac{y}{y^2+\varepsilon^2}\,dy, \end{aligned}$$

so

$$\begin{aligned} Cf(x+i\varepsilon)+Cf(x-i\varepsilon)-\frac{i}{\pi}Hf(x) &= \frac{i}{\pi}\,\text{p.v.}\int_{-\infty}^{\infty} f(x-y)\left\{\frac{y}{(y^2+\varepsilon^2)}-\frac{1}{y}\right\}dy \\ &= -\frac{i}{\pi}\,\text{p.v.}\int_{-\infty}^{\infty}\frac{\varepsilon^2 f(x-y)}{y(y^2+\varepsilon^2)}\,dy. \end{aligned}$$

We may use the same trick as before. The function $y/(y^2+\varepsilon^2)$ is integrable with respect to $y$ and is odd, so the preceding principal value integral is

$$\frac{i}{\pi}\int_{-\infty}^{\infty}\frac{f(x-y)-f(y)}{y}\cdot\frac{\varepsilon^2}{y^2+\varepsilon^2}\,dy.$$

Since $f'$ is bounded, this integral is dominated by

$$\int_{-\infty}^{\infty}\frac{\varepsilon^2}{y^2+\varepsilon^2}\,dy \;=\; \varepsilon\int_{-\infty}^{\infty}\frac{dx}{1+x^2} \;=\; \frac{\varepsilon}{\pi}.$$

Thus (18.2.7) converges uniformly. As in the previous proof, the derivative of (18.2.3) is (18.2.3) with $f'$ in place of $f$. Uniform convergence of (18.2.3) follows from the boundedness of this derivative. □

For $z$ not real,

$$Cf(z) \;=\; \frac{1}{2}\left[Cf(z)-Cf(\bar z)\right]+\frac{1}{2}\left[Cf(z)+Cf(\bar z)\right].$$

Therefore Theorems 18.2.1 and 18.2.2 have the following consequence.

**Corollary 18.2.3.** (Plemelj–Sokhotski formulas) *For $f \in \mathcal{S}$, the function $Cf(z)$ for* $\operatorname{Im} z > 0$ *extends continuously to the closure $\mathbb{C}_+ \cup \mathbb{R}$ with value $C_+f(\xi)$, $\xi \in \mathbb{R}$. Similarly, $Cf(z)$ for* $\operatorname{Im} z < 0$ *extends continuously to the closure $\mathbb{C}_- \cup \mathbb{R}$, with value $C_-f(\xi)$, $\xi \in \mathbb{R}$. These one-sided limiting values are given by*

$$C_+f(x) \;=\; \frac{1}{2}f(x)+\frac{i}{2\pi}Hf(x); \qquad C_-f(x) \;=\; -\frac{1}{2}f(x)+\frac{i}{2\pi}Hf(x). \tag{18.2.8}$$

**Remark.** The condition that $f$ belongs to $\mathcal{S}$ can be weakened considerably. See Exercises 5 and 6 for conditions that are closer to being minimal.

## 18.3 The Fourier transform

Again we assume throughout that the function $f$ belongs to the Schwartz class $\mathcal{S}$. Let us take a second look at the Cauchy transform of $f$. Note that if $y \in \mathbb{R}$ and $\operatorname{Im} z > 0$, then

$$\frac{1}{i(y-z)} = \int_0^\infty e^{i(z-y)\xi}\, d\xi, \tag{18.3.1}$$

while if $\operatorname{Im} z < 0$,

$$\frac{1}{i(y-z)} = -\int_{-\infty}^0 e^{i(z-y)\xi}\, d\xi. \tag{18.3.2}$$

Setting $z = x \pm i\varepsilon$, $\varepsilon > 0$, we have

$$Cf(x+i\varepsilon) = \frac{1}{2\pi}\int_{-\infty}^\infty \int_0^\infty e^{-\varepsilon\xi} e^{i(x-y)\xi} f(y)\, d\xi\, dy$$

and

$$Cf(x-i\varepsilon) = -\frac{1}{2\pi}\int_{-\infty}^\infty \int_{-\infty}^0 e^{\varepsilon\xi} e^{i(x-y)\xi} f(y)\, d\xi\, dy.$$

Therefore our previous result about convergence on vertical lines says that

$$f(x) = \frac{1}{2\pi}\lim_{\varepsilon\to 0+}\int_{-\infty}^\infty \int_{-\infty}^\infty e^{-\varepsilon|\xi|} e^{i(x-y)\xi} f(y)\, d\xi\, dy.$$

Under our assumption on $f$ the integrand is absolutely integrable. We may interchange the order of integration and obtain

$$f(x) = \frac{1}{2\pi}\lim_{\varepsilon\to 0+}\int_{-\infty}^\infty e^{-\varepsilon|\xi|} e^{ix\xi}\left\{\int_{-\infty}^\infty e^{-iy\xi} f(y)\, dy\right\} d\xi. \tag{18.3.3}$$

For now we define the *Fourier transform* of $f$ to be the function

$$\widehat{f}(\xi) = \int_{-\infty}^\infty e^{-ix\xi} f(x)\, dx. \tag{18.3.4}$$

(There are a number of different normalizations for the Fourier transform. They take the form

$$\widehat{f}(\xi) = A\int_{-\infty}^\infty e^{iBx\xi} f(x)\, dx,$$

usually with $B = \pm 1$ or $\pm 2\pi$ and $A = 1$, $A = 1/(2\pi)$, or $A = 1/\sqrt{2\pi}$. In fact we will use a second normalization later in this chapter and a third in Chapter 20.)

As we shall see, the assumption $f \in \mathcal{S}$ implies that $\widehat{f}$ is in $\mathcal{S}$. In particular $|\widehat{f}|$ is integrable and (18.3.3) becomes the *inversion formula*

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ix\xi}\widehat{f}(\xi)\,d\xi. \qquad (18.3.5)$$

The transformation

$$\check{g}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ix\xi}g(\xi)\,d\xi \qquad (18.3.6)$$

is called the *inverse Fourier transform* of $g$.

Our assumption on $f$ implies that the formula (18.3.4) can be differentiated to give

$$[\widehat{f}]'(\xi) = \int_{-\infty}^{\infty} e^{-ix\xi}[-ixf(x)]\,dx. \qquad (18.3.7)$$

Multiplying (18.3.4) by $i\xi$ and integrating by parts gives

$$i\xi\,\widehat{f}(\xi) = \int_{-\infty}^{\infty} e^{-ix\xi}f'(x)\,dx. \qquad (18.3.8)$$

It follows that products of polynomials with $\widehat{f}$ and its derivatives are bounded. The same considerations apply to the inverse transform. Thus we have the following:

**Proposition 18.3.1.** *The Fourier transform, or the inverse Fourier transform, of a Schwartz function is a Schwartz function.*

## 18.4 The Fourier transform for $L^1(\mathbb{R})$

We may consider the space $L^1 = L^1(\mathbb{R})$ to be the completion of $\mathcal{S}$ with respect to the metric

$$d(f,g) = ||f-g||_1 = \int_{-\infty}^{\infty} |f(x)-g(x)|\,dx.$$

The elements of $L^1$ can be considered as equivalence classes of Cauchy sequences $\{f_n\}$ from $\mathcal{S}$. Two such sequences $\{f_n\}$, $\{g_n\}$ are equivalent if $||f_n - g_n|| \to 0$. An element $f \in \mathcal{S}$ can be identified with the constant sequence $f_n = f$, and thus taken to belong to $L^1$. According to Theorem 18.1.2, a function that is constant on a bounded interval and vanishes outside that interval is the limit of such a sequence, so $L^1$ contains every piecewise constant function that vanishes for large $|x|$. Conversely each function in $\mathcal{S}$ can be approximated in the $L^1$ sense by such functions, so such piecewise constant functions are dense in $L^1$.

(Technically the elements of $L^1$ are equivalence classes of integrable Lebesgue-measurable functions, where two such functions are equivalent if they differ only on a set of measure zero. We shall write elements of $L^1$ as though they are functions, but for the purposes of this chapter no knowledge of measure theory is required: every argument can be reduced to a statement about piecewise constant functions or functions in the space $\mathcal{S}$.)

**Proposition 18.4.1.** *The Fourier transform extends to a continuous map from $L^1(\mathbb{R})$ into $C_0(\mathbb{R})$.*

*Proof:* Suppose that $f$ belongs to $\mathcal{S}$. Then

$$|\widehat{f}(\xi)| \leq \int_{-\infty}^{\infty} |e^{-ix\xi} f(x)|\, dx = \int_{-\infty}^{\infty} |f(x)|\, dx = ||f||_1. \tag{18.4.1}$$

Since $\widehat{f}$ is also a Schwartz function, $f$ belongs to $C_0(\mathbb{R})$. Density of $\mathcal{S}$, together with (18.4.1), implies the extension property. □

**Remark.** Not every element of $C_0(\mathbb{R})$ is the Fourier transform of an element of $L^1(\mathbb{R})$: Exercise 7.

The operation of convolution, (18.1.6) clearly carries over to pairs $f$, $g$ of Schwartz class functions. It is not difficult to show that $f * g$ is bounded and

$$[f * g]' = f' * g = f * g',$$

while if $f_1(x) = xf(x)$ and $g_1(x) = xg(x)$, then

$$x[f * g](x) = f_1 * g(x) + f * g_1(x).$$

Iterating, it follows $\mathcal{S}$ is closed under convolution.

**Proposition 18.4.2.** *Convolution extends by continuity to all $f$, $g$ in $L^1(\mathbb{R})$.*

*Proof:* For $f$ and $g$ in $\mathcal{S}$,

$$\begin{aligned} ||f * g||_1 &\leq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |f(x-y)g(y)|\, dy\, dx \\ &= \int_{-\infty}^{\infty} |g(y)| \left\{ \int_{-\infty}^{\infty} |f(x-y)\, dx \right| dy \\ &= \int_{-\infty}^{\infty} |g(y)|\, ||f||_1\, dy = ||f||_1\, ||g||_1. \end{aligned}$$

It follows from this inequality that if $\{f_n\}$ and $\{g_n\}$ are Cauchy sequences from $\mathcal{S}$, then $f_n * g_n$ is a Cauchy sequence. In fact Cauchy sequences are bounded, and

$$\begin{aligned} ||f_n * g_n - f_m * g_m||_1 &= ||f_n * g_n - f_m * g_n + f_m * g_n - f_m * g_m||_1 \\ &\leq ||(f_n - f_m) * g_n||_1 + ||f_m * (g_n - g_m)||_1 \\ &\leq (||f_n - f_m||_1 + ||g_n - g_m||_1) \cdot \sup_k \{||f_k||_1 + ||g_k||_1\}. \end{aligned}$$ □

**Theorem 18.4.3.** *If $\{G_\varepsilon\}$ is an approximate identity and $f$ belongs to $L^1$, then $G_\varepsilon * f$ converges to $f$ in $L^1$:*

$$\lim_{\varepsilon \to 0} ||G_\varepsilon * f - f||_1 = 0. \tag{18.4.2}$$

*Proof:* It is enough to prove this for a dense subset. Suppose that $f \in \mathcal{S}$. Then, in view of (18.1.2),

$$\begin{aligned}
|[G_\varepsilon * f](x) - f(x)| &= \int_{-\infty}^{\infty} |f(x-y) - f(x)| G_\varepsilon(y)\,dy \\
&= \int_{|y|<\delta} |f(x-y) - f(x)| G_\varepsilon(y)\,dy \\
&\quad + \int_{|y|>\delta} |f(x-y) - f(x)| G_\varepsilon(y)\,dy \\
&\leq \delta \sup_x |f'(x)| + 2 \sup_x |f(x)| \int_{|y|>\delta} G_\varepsilon(y)\,dy.
\end{aligned}$$

The first term in the last line is small for small $\delta$, and for fixed $\delta$ the second term is small for small $\varepsilon$. □

The space $L^1(\mathbb{R})$ is a ring with convolution as multiplication. It is an important fact that the Fourier transform is an isomorphism of $L^1(\mathbb{R})$ into the ring $C_0(\mathbb{R})$ of continuous functions with limit 0 as $|x| \to \infty$.

**Theorem 18.4.4.** *If $f$ and $g$ belong to $L^1(\mathbb{R})$, then the Fourier transform of the convolution is the product of the Fourier transforms:*

$$\widehat{f * g}(\xi) \;=\; \widehat{f}(\xi)\,\widehat{g}(\xi). \tag{18.4.3}$$

*Proof:* It is enough to prove this for a dense set of functions. Assume that $f$ and $g$ belong to $\mathcal{S}$. Then

$$\begin{aligned}
\widehat{f * g}(\xi) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-ix\xi} f(x-y)\, g(y)\,dy\,dx \\
&= \int_{-\infty}^{\infty} e^{-iy\xi} g(y) \left\{ \int_{-\infty}^{\infty} e^{-i(x-y)\xi} f(x-y)\,dx \right\} dy \\
&= \int_{-\infty}^{\infty} e^{-iy\xi} g(y)\,[\widehat{f}(\xi)]\,dy \;=\; \widehat{f}(\xi)\,\widehat{g}(\xi). \qquad \square
\end{aligned}$$

## 18.5 The Fourier transform for $L^2(\mathbb{R})$

In working with $L^2$, it is convenient to change the normalization so that the formulas for the transform and the inverse transform become nearly symmetric. Thus we change notation and set, for $f$, $g$ in $\mathcal{S}$,

$$\widehat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-ix\xi} f(x)\,dx; \tag{18.5.1}$$

$$\check{g}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ix\xi} g(\xi)\,d\xi. \tag{18.5.2}$$

We take the $L^2$ *inner product* of two Schwartz class functions $f$ and $g$ to be the integral

$$(f,g) \;=\; \int_{-\infty}^{\infty} f(x)\,\overline{g(x)}\,dx. \tag{18.5.3}$$

The associated $L^2$ *norm* is

$$||f||_2 \;=\; (f,f)^{1/2} \;=\; \left[\int_{-\infty}^{\infty} |f(x)|^2\,dx\right]^{1/2}. \tag{18.5.4}$$

As noted in Section 1.9, there is an associated Cauchy–Schwarz inequality

$$|(f,g)| \;\le\; ||f||_2\,||g||_2. \tag{18.5.5}$$

The space $L^2(\mathbb{R})$ is the completion of $\mathcal{S}$ with respect to the metric induced by the $L^2$ norm. As in the case of $L^1$, various other function spaces are dense in $L^2$, such as piecewise constant functions, or continuous functions, that vanish outside a bounded interval. The inner product and the Cauchy–Schwarz inequality extend to $L^2(\mathbb{R})$.

The space $L^2(\mathbb{R})$ is especially well-suited for the Fourier transform, and conversely.

**Theorem 18.5.1.** (Plancherel theorem) *The Fourier transform and the inverse transform, as normalized in (18.5.1) and (18.5.2), map $L^2(\mathbb{R})$ onto $L^2(\mathbb{R})$ and preserve the inner product and the norm.*

*Proof:* It is sufficient to prove that for any two Schwartz class functions $f$ and $g$, the inner products satisfy

$$(f,g) \;=\; (\widehat{f},\widehat{g}). \tag{18.5.6}$$

But

$$\begin{aligned}(f,g) &= \int_{-\infty}^{\infty}\left\{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{ix\xi}\,\widehat{f}(\xi)\,d\xi\right\}\overline{g}(x)\,dx\\ &= \int_{-\infty}^{\infty}\widehat{f}(\xi)\left\{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\overline{e^{-ix\xi}\,g(x)\,dx}\right\}d\xi\\ &= \int_{-\infty}^{\infty}\widehat{f}(\xi)\,\overline{\widehat{g}(\xi)}\,d\xi \;=\; (\widehat{f},\widehat{g}). \qquad \square\end{aligned}$$

## Exercises

1. Verify (18.1.7), commutativity of convolution.
2. Show that (18.2.5) is an approximate identity.
3. Complete the proof of uniform convergence of (18.2.2) in Theorem 18.2.1.
4. Compute the Cauchy transform of the function $f$, where

$$f(x) = 1, \quad |x| < 1, \qquad f(x) = 0, \quad |x| > 0.$$

5. Show that Theorem 18.2.1 is valid under the assumptions that for some constants $0 < \alpha < 1$, $\beta > 0$, and $C_j > 0$,

$$|f(x) - f(y)| \le C_1 |x-y|^{\alpha}, \qquad |f(x)| \le C_2 (1+|x|)^{-\beta}, \qquad \text{all } x, y \text{ in } \mathbb{R}.$$

(The first inequality here is known as a *Hölder condition*.) .

6. Show that Theorem 18.2.2 is valid under the assumptions of Exercise 5.
7. Show that the function

$$g(\xi) = 1 - \sqrt{|\xi|}, \quad |\xi| < 1, \qquad g(\xi) = 0, \quad |\xi| \ge 1,$$

is not the Fourier transform of a function in $L^1$. (Show that the inverse Fourier transform of $g$ is $\sin x/(\pi x) + O(|x|^{-3/2})$ for large $|x|$.)

8. Suppose that $f$ belongs to $L^1$. Verify the following statements.
   (a) If $f(-x) = f(x)$, then $\widehat{f}(-\xi) = \widehat{f}(\xi)$.
   (b) If $f(-x) = -f(x)$, then $\widehat{f}(-\xi) = -\widehat{f}(\xi)$.
   (c) If $a > 0$, the Fourier transform of $f(ax)$ is $a^{-1}\widehat{f}(\xi/a)$.
   (d) If $a \in \mathbb{R}$, the Fourier transform of $f(x+a)$ is $e^{ixa}\widehat{f}(\xi)$.
   (e) If $a \in \mathbb{R}$, the Fourier transform of $e^{iax} f(x)$ is $\widehat{f}(\xi - a)$.
9. Find the Fourier transforms of $\cos x\, f(x)$ and $\sin x\, f(x)$ in terms of $\widehat{f}$.
10. Compute the Fourier transform of the function $f$ of Exercise 4.
11. Verify the following:
    (a) The Fourier transform of $1/(1+x^2)$ is $\pi e^{-|\xi|}$.
    (b) The Fourier transform of $e^{-|x|}$ is $2/(1+\xi^2)$.
12. Show that the Fourier transform of the Gaussian $G(x) = (2\pi)^{-1/2} e^{-x^2/2}$ is $e^{-\xi^2/2}$ (Hint: complete the square in the exponential, interpret the integral as the integral over a line in the complex plane, and use Cauchy's theorem to bring it back to the real axis.)
13. Show that the Fourier transform of $1/\cosh x$ is $\pi \cosh(\pi\xi/2)$. (Write $1/(1+e^{-2x})$ and $1/(1+e^{2x})$ for $x > 0$ and $x < 0$, respectively, as convergent series, and integrate. Compare with the partial fractions expansion of $\cosh(\pi x/2)$.)
14. Suppose that $f$ belongs to $L^2(\mathbb{R})$. Given $R > 0$, let

$$f_R(x) = \begin{cases} f(x), & |x| < R; \\ 0, & |x| \ge R. \end{cases}$$

    (a) Prove that each $f_R$ belongs to $L^1(\mathbb{R})$.
    (b) Prove that $||f_R - f||_2 \to 0$ as $R \to \infty$.
15. In this exercise we use the $L^2$ normalization (18.5.1), (18.5.2). The mathematical formulation of the *Heisenberg uncertainty principle* establishes a limit on how tightly a function and its Fourier transform can each be concentrated near a

point. Suppose, for convenience, that $f$ is a Schwartz function with $||f||_2 = 1$, so also $||\widehat{f}||_2 = 1$. Then both $|f(x)|^2$ and $|\widehat{f}(\xi)|^2$ can be considered as probability densities. The integrals

$$\int_{-\infty}^{\infty} x^2|f(x)|^2\,dx, \qquad \int_{-\infty}^{\infty} \xi^2|\widehat{f}(\xi)|^2\,d\xi,$$

are measures of how tightly concentrated $f$ and $\widehat{f}$ are near $x = 0$ and near $\xi = 0$, respectively. (The general case, concentrations near $x = a$, $\xi = b$ can be reduced to this case; see Exercise 8 (c), (d).) The one-dimensional version of the uncertainty principle is

$$\left(\int_{-\infty}^{\infty} x^2|f(x)|^2\,dx\right)\cdot\left(\int_{-\infty}^{\infty} \xi^2|\widehat{f}(\xi)|^2\,d\xi\right) \;\ge\; \frac{1}{4\pi}. \tag{18.5.7}$$

Prove (18.5.7). (Hint: Show that

$$1 \;=\; ||f||^2 \;=\; -\left[(xf, f') + (f', xf)\right].$$

Then use the Cauchy–Schwarz inequality and the renormalized version of (18.3.8).

16. Suppose that $f$ in Exercise 15 is the Gaussian density $G$ of Exercise 12. Show that in this case the lower bound $1/4\pi$ in (18.5.7) is attained. (Note that the normalizations of the Fourier transform in Exercises 12 and 15 are different.)
17. Prove the *Poisson summation formula* for Schwartz functions: if $f$ is such a function, then

$$\sum_{n=-\infty}^{\infty} f(n) \;=\; \sum_{m=-\infty}^{\infty} \widehat{f}(2m\pi). \tag{18.5.8}$$

Show first that

$$\sum_{|n|\le N} f(n) \;=\; \frac{1}{2\pi}\int_{-\infty}^{\infty} D_N(\xi)\,\widehat{f}(\xi)\,d\xi, \tag{18.5.9}$$

where $D_N$ is the *Dirichlet kernel*

$$D_N(\xi) \;=\; \sum_{|n|\le N} e^{in\xi} \;=\; \frac{\sin([N+\frac{1}{2}]\xi)}{\sin\frac{1}{2}\xi},$$

and note that $D_N$ is periodic with period $2\pi$, and its integral over the interval $(-\pi,\pi)$ is $2\pi$, so that

$$\begin{aligned}\frac{1}{2\pi}\int_{-\pi}^{\pi} D_N(\xi)g(\xi)\,d\xi - g(0) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} D_N(\xi)[g(\xi)-g(0)]\,d\xi \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \sin([N+\tfrac{1}{2}]\xi)\,h(\xi)\,d\xi\end{aligned} \tag{18.5.10}$$

where

$$h(\xi) \;=\; \frac{g(\xi)-g(0)}{\sin\frac{1}{2}\xi}.$$

Integrate by parts to show that (18.5.10) converges to zero. The next step is to convert the integral in (18.5.9) into a sum of integrals like that in (18.5.10), leading to (18.5.8).

18. Use Exercise 17 and the property of the Fourier transform under translation to prove the more general form

$$\sum_{n=-\infty}^{\infty} f(ax+n) \;=\; \frac{1}{a}\sum_{m=-\infty}^{\infty} \widehat{f}(2m\pi/a).$$

19. Use Exercises 18 and 12 to prove Jacobi's theta function identity (13.6.3).

## Remarks and further reading

For more on the Cauchy transform and its applications, see Bell [17].

Study of the Hilbert transform leads to the study of singular integrals; see Stein [131] and Estrada [43]. The Hilbert transform plays a role in the study of the Beltrami equation $\bar{\partial} f = \mu\partial z$, which was mentioned earlier in connection with conformal mapping. This equation plays many roles in complex analysis and partial differential equations, see Ahlfors and Bers [6] and Astala, Iwaniec, and Martin [13].

The classic treatment of the Fourier transform is Titchmarsh [136]. There are a number of modern textbooks, e.g. Osgood [112].

# Chapter 19
# Theorems of Phragmén–Lindelöf and Paley–Wiener



The various versions of the theorem of Phragmén and Lindelöf are far-reaching extensions, to unbounded domains, of the maximum modulus principle. The basic conclusion is that a function holomorphic on an unbounded domain $\Omega$, and continuous on the closure, either grows very fast at infinity or is bounded by its values on the boundary of $\Omega$. This has a number of interesting applications. Among them are a theorem of Hardy that characterizes the Gaussian probability distribution, and a theorem of Paley and Wiener that characterizes the Fourier transforms of functions that live on a bounded interval. The Paley–Wiener theorem itself was applied by Hardy to derive a representation of certain entire functions in terms of their values on the integers.

Section 19.1 is self-contained. Subsequent sections involve some knowledge of the Fourier transform, as in Section 18.5.

## 19.1 Phragmén–Lindelöf theorems

The function

$$f(z) = e^z, \qquad \operatorname{Re} z \geq 0, \tag{19.1.1}$$

illustrates how badly the maximum modulus principle can fail in an unbounded domain. On the boundary, the imaginary axis, we have $|f(it)| = 1, t \in \mathbb{R}$, but $f$ grows exponentially on the positive real axis. The Phragmén–Lindelöf principle says that this is the only way the maximum modulus principle can fail in a sector: the function must grow rapidly enough in the interior.

We begin with the simplest version of this principle, which shows that the example (19.1.1) is sharp, in terms of growth like the exponential of a power of $|z|$.

**Theorem 19.1.1.** *Suppose that $f$ is holomorphic in the half plane $\Omega = \{z : \operatorname{Re} z > 0\}$ and continuous up to the imaginary axis. Suppose also that*



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_19

$$|f(z)| \leq M, \quad \mathrm{Re}\, z = 0 \tag{19.1.2}$$

*and suppose that for some* $0 < \beta < 1$ *and some constant* $C$*,*

$$|f(z)| \leq C e^{|z|^\beta}, \quad \mathrm{Re}\, z > 0. \tag{19.1.3}$$

*Then*

$$|f(z)| \leq M, \quad \mathrm{Re}\, z > 0. \tag{19.1.4}$$

*Proof:* Choose $\gamma$ such that $\beta < \gamma < 1$. Given $\varepsilon > 0$, let

$$F_\varepsilon(z) = e^{-\varepsilon z^\gamma} f(z), \quad \mathrm{Re}\, z > 0.$$

Then for $z = re^{i\theta}$,

$$|F_\varepsilon(z)| \leq e^{-\varepsilon r^\gamma \cos\gamma\theta} |f(z)|.$$

If $z = re^{i\theta}$ belongs to $\Omega$ then $\cos\gamma\theta \geq \cos\frac{1}{2}\pi\gamma = \delta > 0$. Combining this with (19.1.3) we have

$$|F_\varepsilon(re^{i\theta})| \leq C e^{r^\beta - \varepsilon\delta r^\gamma}.$$

Since $\beta < \gamma$, $|F_\varepsilon(z)| \to 0$ as $|z| \to \infty$, $z \in \Omega$. It follows from the maximum modulus theorem, Theorem 1.2.4, that $|F_\varepsilon|$ attains its maximum on the boundary:

$$|F_\varepsilon(z)| \leq \sup_{\mathrm{Re}\, z=0} |F_\varepsilon(z)| \leq \sup_{r\geq 0} e^{-\delta\varepsilon r^\gamma} M \leq M; \tag{19.1.5}$$

see Exercise 1. The inequality (19.1.5) is true independent of $\varepsilon > 0$. Since $F_\varepsilon(z) \to f(z)$ as $\varepsilon \to 0$, (19.1.5) implies (19.1.4). □

The same argument works on a general sector

$$\Omega_a = \{z : |\arg z| < a\pi\}, \quad 0 < a < 1. \tag{19.1.6}$$

Theorem 19.1.1 is the case $a = 1/2$.

**Theorem 19.1.2.** *Suppose that* $f$ *is holomorphic on the sector* $\Omega_a$ *and continuous up to the boundary. Suppose that* $|f(z)| \leq M$ *for* $z$ *in the boundary, and suppose that for some* $\beta < 1/2a$ *and some constant* $C$*,*

$$|f(z)| \leq C e^{r^\beta}, \quad z \in \Omega_a. \tag{19.1.7}$$

*Then*

$$|f(z)| \leq M, \quad z \in \Omega_a. \tag{19.1.8}$$

The proof is left as Exercise 2

The same sort of idea occurs in the proof of Theorem 23.1.1, which starts by showing that if a function is holomorphic in a strip $\Omega = \{z : a < \mathrm{Re}\, z < b\}$, bounded, and continuous on the closure, then for $z \in \Omega$

$$|f(z)| \le \max\{\sup_{\mathrm{Re}\, z=a} |f(z)|, \sup_{\mathrm{Re}\, z=b} |f(z)|\}.$$

## 19.2 Hardy's uncertainty principle

The Gaussian probability distribution with mean zero and variance one is the function

$$G(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

It has the property that its Fourier transform

$$\widehat{G}(\xi) = \int_{-\infty}^{\infty} e^{-ix\xi} G(x)\, dx$$

is a multiple of itself:

$$\widehat{G}(\xi) = e^{-\xi^2/2}. \tag{19.2.1}$$

This is Exercise 12 of Chapter 18. (The idea is to complete the square in the exponential, view the integration as occurring along a line parallel to the real axis, and move the integration to the real axis via Cauchy's theorem.)

Hardy [57] showed that if a function $f$ has the property that $f$ and its Fourier transform are both of the order of $G$ at infinity, then $f$ must be a multiple of $G$.

**Theorem 19.2.1.** (Hardy) *Suppose that $f : \mathbb{R} \to \mathbb{C}$ is integrable, and suppose that $f(x) = O(e^{-x^2/2})$ as $|\xi| \to \infty$ and $\widehat{f}(\xi) = O(e^{-\xi^2/2})$ as $|\xi| \to \infty$. Then $f(x) = c\, e^{-x^2/2}$ for some constant $c$.*

*Proof:* If the even part of $f$, $f_e(x) = \frac{1}{2}[f(x) + f(-x)]$, and odd part of $f$, $f_o(x) = \frac{1}{2}[f(x) - f(-x)]$, are each $O(e^{-x^2/2})$, then so is $f$. Therefore we may treat the even and odd cases separately.

Suppose first that $f$ is even. Then $\widehat{f}$ is also even. Choose the constant $C_0$ so that for every real $x$, $\xi$,

$$|f(x)| \le C_0\, e^{-x^2/2}, \qquad |\widehat{f}(\xi)| \le C_0\, e^{-\xi^2/2}.$$

The rapid decay condition $O(e^{-x^2/2})$ implies that $\widehat{f}$ extends to an entire function

$$F(z) = \int_{-\infty}^{\infty} f(x)\, e^{-ixz}\, dx.$$

Then

$$
\begin{aligned}
|F(\xi+i\tau)| &\le C_0 \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\, e^{x\tau}\, dx \\
&= C_0 \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-\tau)^2+\frac{1}{2}\tau^2}\, dx \\
&= C_0\sqrt{2\pi}\, e^{\tau^2/2} \;=\; C\, e^{\tau^2/2}.
\end{aligned} \tag{19.2.2}
$$

Since $F$ in this case is even, its Taylor expansion at the origin has only even powers of $z$. Thus $F(z) = \phi(z^2)$, where $\phi$ is also an entire function.

With $z = re^{i\theta}$, we have $\operatorname{Im}\sqrt{z} = \sqrt{r}\sin(\theta/2)$, so

$$
|\phi(re^{i\theta})| \;=\; |F(\sqrt{r}e^{i\theta/2})| \;\le\; C\, e^{r\sin^2(\theta/2)/2} \;\le\; C\, e^{r/2}. \tag{19.2.3}
$$

For $z = r > 0$, since $C_0 < C$ we have

$$
|\phi(r)| \;=\; |F(\sqrt{r})| \;=\; |\widehat{f}(\sqrt{r})| \;\le\; C\, e^{-r/2} \tag{19.2.4}
$$

Choose $\alpha$ with $0 < \alpha < \pi/2$ and let

$$
w(z,\alpha) \;=\; w(re^{i\theta},\alpha) \;=\; \exp\left(iz\frac{e^{-i\alpha}}{2\sin\alpha}\right) \;=\; \exp\left(ir\frac{e^{i(\theta-\alpha)}}{2\sin\alpha}\right).
$$

Now

$$
\begin{aligned}
ir\frac{e^{i(\theta-\alpha)}}{2\sin\alpha} &= \frac{ir}{2\sin\alpha}\,[\cos(\theta-\alpha)+i\sin(\theta-\alpha)] \\
&= -\frac{r\sin(\theta-\alpha)}{2\sin\alpha} + i\frac{r\cos(\theta-\alpha)}{2\sin\alpha}.
\end{aligned}
$$

Therefore

$$
|w(re^{i\theta},\alpha)| \;=\; \exp\left(-\frac{r\sin(\theta-\alpha)}{2\sin\alpha}\right). \tag{19.2.5}
$$

In particular

$$
|w(r,\alpha)| \;=\; e^{r/2}, \qquad |w(re^{2i\alpha},\alpha)| \;=\; e^{-r/2}. \tag{19.2.6}
$$

Combining (19.2.5) with (19.2.3), we obtain

$$
|w(re^{i\theta},\alpha)\phi(re^{i\theta})| \;\le C\exp\left(\frac{r}{2}\left[1+\frac{1}{\sin\alpha}\right]\right). \tag{19.2.7}
$$

Combining (19.2.3) and (19.2.4) with (19.2.5), we obtain

$$
|w(r,\alpha)\,\phi(r)| \;\le\; C, \qquad |w(re^{2i\alpha},\alpha)\,\phi(re^{2i\alpha})| \;\le\; C. \tag{19.2.8}
$$

The Phragmén–Lindelöf principle applies, giving

$$
|w(z,\alpha)\,\phi(z)| \;\le\; C \quad \text{for } \; 0 \le \arg z \le 2\alpha; \tag{19.2.9}
$$

see Exercise 6. Therefore in this sector

$$|\phi(re^{i\theta})| \le C\exp\left(\frac{r\sin(\theta-\alpha)}{2\sin\alpha}\right).$$

For fixed $\theta$, let $\alpha \to \pi/2$ to conclude that

$$|\phi(z)| \le Ce^{-r\cos\theta/2}, \quad 0 \le \arg z < \pi. \tag{19.2.10}$$

By continuity, this also holds for $\arg z = \pi$.

The half plane $\mathbb{C}_-$ can be treated in a similar way, leading to the same estimate (19.2.10). It follows that for each $z = re^{i\theta}$,

$$|e^{z/2}\phi(z)| = e^{r\cos\theta/2}|\phi(z)| \le C.$$

Therefore the entire function $e^{z/2}\phi(z)$ is a constant $c$. This implies that $F(z) = ce^{-z^2/2}$. In particular, the Fourier transform $\widehat{f}(\xi)$ is $ce^{-\xi^2/2}$. This in turn implies that $f$ is $(c/\sqrt{2\pi})e^{-x^2/2}$.

Suppose now that $f$ is odd. Then $F$ is odd, so $F(\xi)/\xi$ is an entire even function that is $O(e^{-\xi^2/2})$. Therefore, by the previous argument, $F(\xi)/\xi$ is a constant multiple of $e^{-\xi^2/2}$. Then $e^{\xi^2/2}F(\xi)/\xi$ is entire and is $O(1/|\xi|)$, so $F = 0$ and $f = 0$. □

**Corollary 19.2.2.** *Suppose that $f : \mathbb{R} \to \mathbb{C}$ is integrable, and suppose that $f(x) = O(e^{-ax^2/2})$ as $|x| \to \infty$ and $\widehat{f}(\xi) = O(e^{-b\xi^2/2})$ as $|\xi| \to \infty$, where $a$ and $b$ are positive. If $ab > 1$ then $f = 0$.*

The proof is left as Exercise 7.

Theorem 19.2.1 and Corollary 19.2.2 are known as *Hardy's uncertainty principle*. Like the Heisenberg uncertainty principle, Hardy's principle has to do with the extent that a function and its Fourier transform can both tail off rapidly, and thus can simultaneously be concentrated on bounded sets.

. .

## 19.3 The Paley–Wiener Theorem

The Fourier transform of a function that vanishes outside a bounded interval is an entire function. The theorem of Paley and Wiener [115] characterizes such transforms. The Fourier transform was defined in Chapter 18 for functions in $L^1(\mathbb{R})$. Suppose that $f$ belongs to $L^1(R)$ and that $f(x) = 0$ for $|x| > A$. The Fourier transform $\widehat{f}$ extends to an entire function

$$F(z) = \int_{-A}^{A} f(x)e^{-ixz}\,dx \tag{19.3.1}$$

that satisfies the inequality

$$|F(z)| \leq \int_{-A}^{A} |f(x)| e^{A|z|}\, dx = O(e^{A|z|}). \tag{19.3.2}$$

Such an entire function $F$ is said to be of *exponential type*. The greatest lower bound of the constants $A > 0$ such that (19.3.2) is valid is said to be the *type* of $F$. The space of entire functions of type at most $A$, i.e. functions $F$ such that

$$|F(z)| = O(e^{(|A|+\varepsilon)|z|}), \tag{19.3.3}$$

for every $\varepsilon > 0$, is denoted $E^A$.

The theorem of Paley and Wiener involves the extension of the Fourier transform to functions $f$ in $L^2(\mathbb{R})$. Suppose that $f$ is such a function, and suppose that $f(x) = 0$ for $|x| > R$. Then the Cauchy–Schwarz inequality shows that $f$ belongs to $L^1(\mathbb{R})$:

$$\int_{-\infty}^{\infty} |f(x)|\, dx = \int_{-R}^{R} |f(x)|\, dx \leq \left[\int_{-R}^{R} |f(x)|^2\, dx \cdot \int_{-R}^{R} 1\, dx\right]^{1/2} = \sqrt{2R}\, ||f||_2.$$

As noted in Section 18.5, in the context of $L^2(\mathbb{R})$, it is convenient to choose $1/\sqrt{2\pi}$ as the normalizing constant for the Fourier transform. Then the transform and its inverse are the same, up to a sign in the exponential:

$$\widehat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x)\, e^{-ix\xi}\, dx; \qquad \check{g}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(\xi)\, e^{ix\xi}\, d\xi.$$

Moreover the inner products and norms are related by

$$(f,g) \equiv \int_{-\infty}^{\infty} f(x)\, \overline{g(x)}\, dx = \int_{-\infty}^{\infty} \widehat{f}(\xi)\, \overline{\widehat{g}(\xi)}\, d\xi \equiv (\widehat{f},\widehat{g}). \tag{19.3.4}$$

Therefore

$$||f||_2 = ||\widehat{f}||_2 \tag{19.3.5}$$

**Theorem 19.3.1.** (Paley–Wiener) *A function $F : \mathbb{R} \to \mathbb{C}$ is the Fourier transform of a function $f \in L^2(\mathbb{R})$ that vanishes for $|x| > A > 0$, i.e.*

$$F(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-A}^{A} f(x) e^{-ix\xi}\, dx, \tag{19.3.6}$$

*if and only if $F$ belongs to $L^2(\mathbb{R})$ and extends to an entire function that belongs to $E^A$.*

*Proof:* We have shown that if $F$ has the form (19.3.6) with $f \in L^2(\mathbb{R})$, then $F$ extends and belongs to $E^A$. By (19.3.5), the restriction to $\mathbb{R}$ belongs to $L^2(\mathbb{R})$.

Conversely, suppose that a function $F$ belongs to $E^A$, and its restriction to $\mathbb{R}$ belongs to $L^2$. Then the inverse transform

$$f(x) = \lim_{R\to\infty} \left\{ \frac{1}{\sqrt{2\pi}} \int_{-R}^{R} F(\xi) e^{ix\xi}\, d\xi \right\} \tag{19.3.7}$$

belongs to $L^2(\mathbb{R})$; see Exercise 14 of Chapter 18. Our goal is to show that $f(x) = 0$ for $|x| > A$.

Let

$$g(z) \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} F(u-z)du. \tag{19.3.8}$$

Then $g(z)$ is an entire function of $z$. Since $F \in E^A$, for $\varepsilon > 0$ we have

$$\begin{aligned} |g(z)| &\le \int_{-\frac{1}{2}}^{\frac{1}{2}} |F(u-z)|du = O\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} e^{(A+\varepsilon)(|z|+|u|)}du\right) \\ &= O\left(e^{(A+\varepsilon)|z|}\int_{-\frac{1}{2}}^{\frac{1}{2}} e^{(A+\varepsilon)|u|}du\right) \;=\; O\left(e^{(A+\varepsilon)|z|}\right). \end{aligned} \tag{19.3.9}$$

Furthermore, for real $x$

$$|g(x)|^2 \;\le\; \int_{-\frac{1}{2}}^{\frac{1}{2}} |F(u-x)|^2du \;\le\; \|F\|_2^2 \;<\; \infty \tag{19.3.10}$$

and

$$\begin{aligned} \int_{-\infty}^{\infty} |g(x)|^2dx &\le \int_{-\infty}^{\infty} \left\{\int_{-\frac{1}{2}}^{\frac{1}{2}} |F(u-x)|^2du\right\}dx \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\infty}^{\infty} |F(u-x)|^2dxdu \\ &\le \int_{-\frac{1}{2}}^{\frac{1}{2}} \|F\|_2^2\, du = \|F\|_2^2 < \infty, \end{aligned} \tag{19.3.11}$$

so that $g(x)$ is bounded and $g(x)$ belongs to $L^2(\mathbb{R})$.

Choose $B > A$ and define

$$G(z) \;=\; e^{iBz}g(z). \tag{19.3.12}$$

Since $g(z)$ belongs to $E^A$, it follows that $G$ is of exponential type. For $z = x$ real, (19.3.10) implies that

$$|G(x)| \;=\; |g(x)| \;=\; O(1). \tag{19.3.13}$$

If $z = ib$ and $b > 0$, then by (19.3.9),

$$|G(ib)| \;=\; |e^{-Bb}g(ib)| \;=\; O(e^{-Bb+(A+\varepsilon)b}).$$

Since $B > A$, for small $\varepsilon > 0$ the quantity on the right side approaches zero as $b \to \infty$. In particular we obtain

$$|G(ib)| \;=\; O(1), \qquad \text{as } b \to \infty. \tag{19.3.14}$$

It follows from (19.3.13), (19.3.14), and the Phragmén–Lindelöf principle, that $|e^{iBz}g(z)|$ is $O(1)$ for $z = Re^{i\theta}, 0 \le \theta \le \pi/2$. Thus

$$|g(Re^{i\theta})| \le c_1 e^{BR\sin\theta} \quad \text{for } 0 \le \theta \le \pi/2.$$

Similarly,

$$|g(Re^{i\theta})| \le c_2 e^{BR\sin\theta} \quad \text{for } \pi/2 < \theta \le \pi.$$

Therefore

$$|g(Re^{i\theta})| = O(e^{BR\sin\theta}), \qquad R > 0, \qquad 0 \le \theta \le \pi. \tag{19.3.15}$$

Take $L > 0$ and $x < -B$ and consider the integral

$$\int_{-R}^{R} \frac{e^{-ixu}}{1 - iLu} g(u)\, du, \tag{19.3.16}$$

where $R$ is large enough that $LR > 1$. By Cauchy's theorem, we may change the path of integration in (19.3.16) to a semicircle. In fact the integral of

$$\frac{e^{-ixu} g(u)}{1 - iLu}$$

over the contour $\Gamma$ in Figure 19.1 vanishes, so

$$\begin{aligned}\int_{-R}^{R} \frac{e^{-ixu}}{1 - iLu} g(u) du &= -\int_{|u|=R, \operatorname{Im} u > 0} \frac{e^{-ixu}}{1 - iLu} g(u) du \\ &= -i \int_0^{\pi} \frac{e^{-ixRe^{i\theta}}}{1 - iLRe^{i\theta}} g(Re^{i\theta}) Re^{i\theta}\, d\theta.\end{aligned}$$

A simple estimation gives

$$\left| \int_{-R}^{R} \frac{e^{-ixu}}{1 - iLu} g(u)\, du \right| \le \frac{CR}{LR - 1} \int_0^{\pi} e^{xR\sin\theta + BR\sin\theta}\, d\theta. \tag{19.3.17}$$

Since we have chosen $LR > 1$ and $x < -B$, while $\sin\theta \ge 0$ for $0 \le \theta \le \pi$, the right side of (19.3.17) approaches zero as $R \to \infty$, and

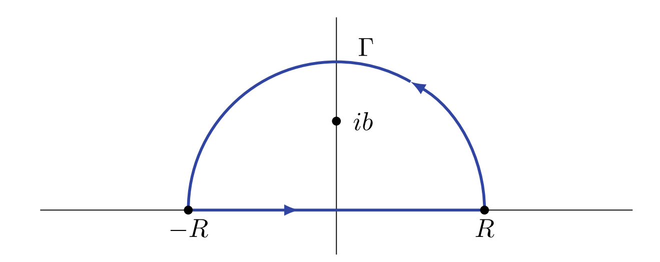



**Fig. 19.1** Change of contour

$$\int_0^\pi e^{-aR\sin\theta}\,d\theta \;=\; O\left(\frac{1}{R}\right) \qquad \text{for every } \; a>0; \tag{19.3.18}$$

see Exercise 8.

Let

$$g_{-L}(u) \;=\; \frac{g(u)}{1-iLu}, \qquad L>0, \qquad -\infty<u<\infty. \tag{19.3.19}$$

Since this is the product of two $L^2$ functions, the Cauchy–Schwarz inequality implies that it belongs to $L^1$. It follows from (19.3.17) and (19.3.18) that the Fourier transform of $g_{-L}$ vanishes for $L>0$ and $x<-B$:

$$\widehat{g}_{-L}(x) \;=\; 0, \qquad L>0,\;\; x<-B. \tag{19.3.20}$$

However (19.3.10) implies that

$$\lim_{L\to 0+}\int_{-\infty}^{\infty}\left|\frac{g(u)}{1-iLu}-g(u)\right|^2 du \;=\; 0. \tag{19.3.21}$$

Hence, by (19.3.5),

$$\lim_{L\to 0+}\int_{-\infty}^{\infty}|\widehat{g}_{-L}(u)-\widehat{g}(u)|^2\,du \;=\; 0. \tag{19.3.22}$$

From (19.3.20), it follows that

$$\widehat{g}(x) \;=\; 0, \qquad \text{for } \; x<-B. \tag{19.3.23}$$

For real $v$

$$\begin{aligned}
g(v) &= \int_{-1/2}^{1/2} F(u-v)du \\
&= \frac{1}{\sqrt{2\pi}}\lim_{R\to\infty}\int_{-1/2}^{1/2}\int_{-R}^{R} f(x)\,e^{-ix(u-v)}\,dx\,du \\
&= \frac{1}{\sqrt{2\pi}}\lim_{R\to\infty}\int_{-R}^{R}\left\{\int_{-1/2}^{1/2} e^{-ixu}\,du\right\} f(x)e^{ixv}\,dx \\
&= \frac{1}{\sqrt{2\pi}}\lim_{R\to\infty}\int_{-R}^{R} f(x)\,\frac{\sin(x/2)}{x/2}\,e^{ixv}\,dx.
\end{aligned}$$

It follows that

$$\widehat{g}(x) \;=\; \sqrt{2\pi}\,f(x)\,\frac{\sin(x/2)}{x/2}.$$

Since (19.3.23) holds for every $B>A$, we conclude that $f(x)=0$ for $x<-A$.

Similarly, we can show that $f(x)=0$ for $x>A$. Consider $G(-ib), b>0$ instead of $G(ib)$, and note that by (19.3.9),

$$|G(-ib)| \;=\; |e^{Bb}g(-ib)| = O(e^{b(A+\varepsilon+B)}), \qquad b>0, \tag{19.3.24}$$

for every $\varepsilon > 0$. If $B < -A < 0$, then for sufficiently small $\varepsilon$, we have

$$|G(-ib)| = O(1) \qquad \text{as } b \to +\infty, \qquad B < -A < 0, \tag{19.3.25}$$

corresponding to (19.3.14). By the Phragmén–Lindelöf principle applied to a semi-circular domain in the lower half plane, we obtain

$$|g(Re^{i\theta})| = O(e^{BR\sin\theta}), \qquad -\pi \le \theta \le 0, \quad R > 0, \tag{19.3.26}$$

corresponding to (19.3.15). Furthermore, for some constant $C$

$$\begin{aligned}\left|\int_{-R}^{R} \frac{e^{-ixu}}{1+iLu} g(u)du\right| &\le \left|\int_{-\pi}^{0} \frac{e^{-ixRe^{i\theta}} g(Re^{i\theta})Re^{i\theta}}{1+iLRe^{i\theta}} d\theta\right| \\ &\le \frac{CR}{LR-1} \int_{-\pi}^{0} e^{(x+B)R\sin\theta} d\theta \\ &= \frac{CR}{LR-1} \int_{0}^{\pi} e^{-(x+B)R\sin\theta} d\theta,\end{aligned}$$

which tends to zero as $R \to \infty$ if $x > -B$. This corresponds to (19.3.17). As in (19.3.20), we deduce that

$$\widehat{g}_L(x) = 0, \qquad L > 0, \qquad x > -B > A, \tag{19.3.27}$$

where $g_L(x) = g(x)/(1+iLx)$ belongs to $L^1(\mathbb{R})$. It follows as before that $f(x) = 0$ for $x > A$. $\square$

**Corollary 19.3.2.** *If $F$ is an entire function of exponential type, and the restriction $f$ of $F$ to $\mathbb{R}$ belongs to $L^2(\mathbb{R})$, then $f(x) \to 0$ as $|x| \to \infty$.*

*Proof:* By Theorem 19.3.1, the inverse Fourier transform of $f$ vanishes outside a bounded interval. Therefore it belongs not only to $L^2(\mathbb{R})$, but to $L^1(\mathbb{R})$. Since $f$ is the Fourier transform of an $L^1$ function, it has limit 0 at $\infty$. $\square$

## 19.4 An application

As a simple application of the Paley–Wiener theorem, together with the basics of Fourier series, we can find formulas that express certain entire functions in terms of their values at the integers. Looked at another way, these are formulas that interpolate from functions on the integers to entire functions.

If $f(z)$ is an entire function of $z$ belonging to the class $E^\sigma, \sigma > 0$, as defined in Section 19.3 then $f\left(\frac{\pi z}{\sigma}\right)$ belongs to $E^\pi$. Thus the study of functions of exponential type may be reduced to that of functions in $E^\pi$.

The basic idea is that if the restriction to $\mathbb{R}$ of such a function $f$ belongs to $L^2(\mathbb{R})$, then the inverse Fourier transform $g$ belongs to $L^2$ and vanishes outside the interval

$[-\pi,\pi]$. The set of such functions is the Hilbert space $H$ with inner product

$$(g,h) = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(x)\overline{h(x)}\,dx;$$

see Section 1.9. A calculation shows that the functions $e^{inx}$, $n = 0, \pm 1, \pm 2, \dots$ are an orthonormal set in $H$. In fact they are a basis; see Section 4.4. Therefore the inverse Fourier transform $g$ of $f$ can be written as a Fourier series:

$$g(x) = \sum_{n=-\infty}^{\infty} a_n e^{inx}, \qquad |x| \le \pi. \tag{19.4.1}$$

where

$$a_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(x)e^{-inx}\,dx = \frac{1}{\sqrt{2\pi}}f(n); \tag{19.4.2}$$

see Theorem 4.4.3.

Since $g$ vanishes outside the interval $[-\pi,\pi]$, we may extend (19.4.1) to all values of $x$ as

$$g(x) = \sum_{n=-\infty}^{\infty} a_n\,\varphi_n(x) \tag{19.4.3}$$

with

$$\varphi_n(x) = \begin{cases} e^{inx}, & |x| \le \pi; \\ 0, & |x| > \pi. \end{cases}$$

The Fourier transform of $\varphi_n$ is

$$\begin{aligned}\widehat{\varphi}_n(\xi) &= \frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi} e^{ix(n-\xi)}\,dx \\ &= \frac{1}{\sqrt{2\pi}}\frac{2\sin\pi(\xi-n)}{(\xi-n)} = (-1)^n\sqrt{\frac{2}{\pi}}\,\frac{\sin\pi\xi}{\xi-n}.\end{aligned} \tag{19.4.4}$$

Combining (19.4.2), (19.4.3), and (19.4.4), we get the following result of Hardy [58].

**Theorem 19.4.1.** (Hardy) *Suppose that $f(z)$ belongs to $E^\pi$ and its restriction to $\mathbb{R}$ belongs to $L^2(\mathbb{R})$. Then*

$$f(z) = \frac{\sin\pi z}{\pi}\sum_{n=-\infty}^{\infty}(-1)^n\frac{f(n)}{z-n}. \tag{19.4.5}$$

This result can easily be adapted to the case when $f$ restricted to $\mathbb{R}$ is only assumed to be bounded.

**Theorem 19.4.2.** *Suppose that $f(z)$ belongs to $E^\pi$ and its restriction to $\mathbb{R}$ is bounded. Then*

$$f(z) = \frac{\sin \pi z}{\pi} \left\{ f'(0) + \frac{f(0)}{z} + \sum_{n \neq 0} (-1)^n f(n) \left[ \frac{1}{z-n} + \frac{1}{n} \right] \right\}. \tag{19.4.6}$$

*Proof:* Put $g(z) = [f(z) - f(0)]/z$. Then $g(z)$ belongs to $E^\pi$, and $g(x)$ is in $L^2(\mathbb{R})$. By Theorem 19.4.1,

$$g(z) = \frac{\sin \pi z}{\pi} \sum_{n=-\infty}^{\infty} (-1)^n \frac{g(n)}{z-n}.$$

Since $g(0) = f'(0)$, it follows that

$$f(z) - f(0) = \frac{z \sin \pi z}{\pi} \sum_{n \neq 0} (-1)^n \frac{f(n) - f(0)}{n(z-n)} + f'(0) \frac{\sin \pi z}{\pi}. \tag{19.4.7}$$

The expansion

$$\frac{\pi}{\sin \pi z} = \frac{1}{z} + \sum_{n \neq 0} (-1)^n \left( \frac{1}{z-n} + \frac{1}{n} \right) = \frac{1}{z} + z \sum_{n \neq 0} \frac{(-1)^n}{n(z-n)} \tag{19.4.8}$$

(see Exercise 10) gives

$$f(0) = f(0) \frac{\sin \pi z}{\pi} \cdot \frac{\pi}{\sin \pi z} = f(0) \frac{\sin \pi z}{\pi} \left( \frac{1}{z} + z \sum_{n \neq 0} \frac{(-1)^n}{n(z-n)} \right).$$

Combining this with (19.4.7) yields (19.4.6). □

## Exercises

1. Suppose that $f$ is holomorphic in $\{z : \operatorname{Re} z > 0\}$, continuous up to the imaginary axis, and has limit 0 as $|z| \to \infty$, $\operatorname{Re} z \geq 0$. Use the maximum modulus principle to show (without using Theorem 19.1.1) that

$$|f(z)| \leq \sup_{t \in \mathbb{R}} |f(it)|.$$

   Show that the inequality is strict unless $f(z) \equiv 0$.
2. Prove Theorem 19.1.2.
3. Show that the restriction on the exponent $\beta$ in Theorems 19.1.1 and Theorem 19.1.2 cannot be relaxed.
4. Suppose that $f$ is entire and $|f(z)| \leq C \exp(|z|^\rho)$, where $\rho < 1/2$. Show that if $f$ is bounded on a ray, then it is constant.
5. Suppose that $\alpha > 1/2$ and $f$ is holomorphic in the domain $\Omega = \{z : |\operatorname{Re} z| < \pi/2\alpha,\ \operatorname{Im} z \geq 0\}$ and continuous on the closure. Suppose also that $f$ is bounded on the vertical sides of the closure, and that $\log |f(s+it)| = O(e^{\beta t})$ as $t \to \infty$, $|s| < \pi/2\alpha$, where $\beta < \alpha$. Prove that $f$ is bounded in $\Omega$.

6. Adapt Theorem 19.1.2 to show that the inequalities (19.2.7) and (19.2.8) imply (19.2.9).
7. (a) Prove Corollary 19.2.2. (Hint: use part (c) of Exercise 8 of Chapter 18.)

   (b) What can be said if $ab = 1$?
8. Prove (19.3.18).
9. Verify directly that the functions $\widehat{\varphi}_n$ in (19.4.4) belong to $E^{\pi}$.
10. Verify the expansion (19.4.8).

## Remarks and further reading

For some history of the Phragmén–Lindelöf principle, see Gårding [47]. The Phragmén–Lindelöf principle occurs again in complex interpolation theory and the proof of the Riesz–Thorin theorem; see Chapter 23.

The Paley–Wiener theorem has applications to the theory of entire functions of exponential type; see Chapter 8 and the references there.

# Chapter 20
# Theorems of Wiener and Lévy; the Wiener–Hopf method



This chapter follows on Sections 18.3 and 18.4. We want to apply the results on the Fourier transform to deal with convolution equations —equations of the form

$$u(x) = \int_{-\infty}^{\infty} k(x-y)\,u(y)\,dy + f(x), \qquad x \in \mathbb{R}, \tag{20.0.1}$$

or the form

$$u(x) = \int_{0}^{\infty} k(x-y)\,u(y)\,dy + f(x), \qquad x > 0. \tag{20.0.2}$$

We assume that $k$ and $f$ belong to $L^1 = L^1(\mathbb{R})$, and look for solutions $u \in L^1$. Wiener proved that if a function $1 - \widehat{k}(\xi)$ does not vanish for any $\xi \in \mathbb{R}$, then its reciprocal has the same form. In view of results from Chapter 18, this result is key to understanding the equation (20.0.1). Wiener's result was generalized by Lévy to general holomorphic functions of $\hat{k}$. An important application is to the subtler case of convolution equations of the form (20.0.2).

The theorems of Wiener and of Lévy are proved in Section 20.1. In Section 20.2 we study equations (20.0.2) by a modified version of the Wiener–Hopf technique, due to Gohberg and Krein.

## 20.1 The ring $\mathscr{R}$

As constructed in Section 18.4, the space $L^1 = L^1(\mathbb{R})$ can be taken to be the completion of the space of continuous or piecewise constant functions $f : \mathbb{R} \to \mathbb{C}$ that vanish for large $|x|$, with respect to the metric induced by the norm

$$||f||_1 = \int_{-\infty}^{\infty} |f(x)|\,dx.$$

$L^1$ can also be taken to be the completion of the space $\mathscr{S}$ of Schwartz functions: functions each of whose derivatives is $O(|x|^{-n})$ as $|x| \to \infty$, all $n$. We have shown

 
R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_20

that the Fourier transform of a Schwartz function is a Schwartz function (Proposition 18.3.1.)

It was also shown that $L^1(\mathbb{R})$ is a commutative ring, with convolution as multiplication:

$$[f*g](x) \;=\; \int_{-\infty}^{\infty} f(x-y)\,g(y)\,dy.$$

It is convenient to add an identity element $\mathbf{1}$ to $L^1$. Thus we consider formal expressions $a\mathbf{1}+f$, $a\in\mathbb{C}$, $f\in L^1$, with the multiplication

$$(a\mathbf{1}+f)*(b\mathbf{1}+g) \;=\; ab\mathbf{1}+ag+bf+f*g.$$

The enlarged ring is denoted $\mathbb{C}\mathbf{1}\oplus L^1$.

The Fourier transform $\widehat{\mathbf{1}}$ is defined to be identically 1. Thus the image of $\mathbb{C}\mathbf{1}\oplus L^1$ under the Fourier transform is a subspace $\mathscr{R}$ of the space of continuous functions $F:\mathbb{R}\to\mathbb{C}$ that have a limit at $\infty$:

$$\lim_{|x|\to\infty} F(x) \;\text{ exists and is finite.}$$

Let $\mathscr{R}_0=\{\widehat{f}: f\in L^1\}$, so $\mathscr{R}=\mathbb{C}\oplus\mathscr{R}_0$. We define a norm in $\mathscr{R}$ by

$$||a+\widehat{f}|| \;=\; |a|+||f||_1, \quad a\in\mathbb{C},\;\; f\in L^1.$$

Since $L^1$ is complete with respect to the metric induced by its norm, so is $\mathscr{R}$.

The Fourier transform takes convolution to multiplication:

$$\widehat{f*g}(\xi) \;=\; \widehat{f}(\xi)\,\widehat{g}(\xi);$$

Theorem 18.4.4. Therefore $\mathscr{R}$ is a ring, and the Fourier transform is a ring isomorphism from $L^1(\mathbb{R})\oplus\mathbb{C}$ to $\mathscr{R}$. This is the key fact in the study of certain types of integral equations.

Not every continuous function from $\mathbb{R}$ to $\mathbb{C}$ with limit zero at $\infty$ belongs to $\mathscr{R}_0$, so not every continuous function from $\mathbb{R}$ to $\mathbb{C}$ with limit at $\infty$ belongs to $\mathscr{R}$; see Exercise 7 of Chapter 18. Therefore the next proposition and the following two theorems are significant.

**Proposition 20.1.1.** *$\mathscr{R}$ contains every rational function that has no poles on the real axis and is regular at $\infty$.*

*Proof:* The identities (18.3.1) and (18.3.2) show that each such function with a single pole belongs to $\mathscr{R}$. These functions generate the subring of rational functions described in the statement. □

**Theorem 20.1.2.** (Wiener) *Suppose that $F$ belongs to $\mathscr{R}$ and* $\lim_{|x|\to\infty}F(x)\neq 0$. *Then $F$ has an inverse in $\mathscr{R}$ if and only if $F(\xi)\neq 0$, all $\xi\in\mathbb{R}$.*

For the statement and proof of the analogous theorem for Fourier series, see Exercises 14 and 15 of Chapter 4. For an adaptation that proves Theorem 20.1.2; see Exercise 5. Theorem 20.1.2 is a special case of the following theorem of Lévy [90].

**Theorem 20.1.3.** *Suppose that $F$ belongs to $\mathscr{R}$ and suppose that $\phi$ is holomorphic on a domain $\Omega$ that contains the closure of the image $F(\mathbb{R})$. Then the composition $\phi \circ F$ belongs to $\mathscr{R}$.*

*Proof:* Let $A \subset \mathbb{C}$ be the closure of $F(\mathbb{R})$. Choose $\delta > 0$ small enough that $\Omega$ contains the $2\delta$ neighborhood of $A$:

$$\{z : |z-a| < 2\delta, \text{ some } a \in A\}.$$

Suppose that $F = c + \widehat{f}$. Choose a Schwartz function $g \in L^1$ such that

$$||f-g|| < \delta/2. \tag{20.1.1}$$

Let $G = c + \widehat{g}$, so that $||F-G|| < \delta/2$. For each $\xi \in \mathbb{R}$, the Cauchy integral formula gives

$$\phi(F(\xi)) = \frac{1}{2\pi i}\int_{|\zeta|=\delta} \frac{\phi(G(\xi)+\zeta)}{G(\xi)+\zeta-F(\xi)}\, d\zeta. \tag{20.1.2}$$

For each fixed $\zeta$ here, the integrand is a function $H_\zeta$ of $\xi$. We want to show that this function belongs to $\mathscr{R}$, uniformly with respect to $\zeta$. First,

$$\frac{1}{G(\xi)+\zeta-F(\xi)} = \frac{1}{\zeta-[F(\xi)-G(\xi)]} = \frac{1}{\zeta}\sum_{n=0}^{\infty}\left(\frac{F(\xi)-G(\xi)}{\zeta}\right)^n.$$

The summands have norm $||F-G||/\delta < 1/2$ so the series converges in $\mathscr{R}$:

$$\frac{1}{G+\zeta-F} \in \mathscr{R}, \quad \left|\left|\frac{1}{G+\zeta-F}\right|\right| < \frac{1}{\delta}\sum_{n=0}^{\infty}\frac{1}{2^n} = \frac{2}{\delta}. \tag{20.1.3}$$

As elements of $\mathscr{R}$, the summands depend continuously on $\zeta$. Convergence of the sum is uniform with respect to $\zeta$ in the circle $\{\zeta : |\zeta| = \delta\}$. Therefore the map $\zeta \to (\zeta + G - F)^{-1}$ is continuous from the circle to $\mathscr{R}$.

To complete the proof, it is enough to obtain the analogous result for the numerator of the integrand in (20.1.2). This will show that the integrand $H_\zeta$ itself is such a function with bound $O(\delta^{-1})$, so the result of the integration in (20.1.2) will belong to $\mathscr{R}$. Now for each $\zeta$,

$$\phi(G(\xi)+\zeta) - \phi(c+\zeta) = \phi(c+\zeta+\widehat{g}(\xi)) - \phi(c+\zeta) = O(\widehat{g}(\xi)) \tag{20.1.4}$$

as $|\xi| \to \infty$. Since $\widehat{g}$ is a Schwartz function, this difference is $O(|\xi|^{-m})$ as $|\xi| \to \infty$, for every $m > 0$. Similar estimates hold for the derivatives of $\phi(G(\xi)+\zeta) - \phi(c+\zeta)$, so this difference is itself a Schwartz function as a function of $\xi$. By Proposition

18.3.1 the difference (20.1.4) is the Fourier transform of a Schwartz function $g_\zeta$. The dependence on $\zeta$ is continuous, so the proof is complete. □

## 20.2 Convolution equations

One type of integral equation that arises in applications has the form

$$u(x) = \int_{-\infty}^{\infty} k(x-y)\,u(y)\,dy + f(x). \tag{20.2.1}$$

Here $k$ and $f$ are given functions that belong to $L^1$, and a solution $u$ is sought in $L^1$. In view of our prior discussion, it is natural to take the Fourier transform:

$$\widehat{u}(\xi) = \widehat{k}(\xi)\,\widehat{u}(\xi) + \widehat{f}(\xi).$$

If $\widehat{k}(\xi) = 1$ for some $\xi_0$, but $\widehat{f}(\xi_0) \neq 0$, there is clearly a problem. Otherwise, by Theorem 20.1.2, $1-\widehat{k}$ has an inverse $1-\widehat{k}_1 \in \mathscr{R}$, and the (unique) solution to (20.2.1) that belongs to $L^1$ is

$$u(x) = f(x) - [k_1 * f](x). \tag{20.2.2}$$

**Remark.** Given $k \in L^1$, the map $u \to k*u$ is a bounded map in each $L^p$ space on the line, $1 \le p < \infty$, and also in the space of bounded continuous functions on the line. Therefore the solution (20.2.2) works for $f$, $u$ in each such space.

A subtler type of integral equation arises in certain physical problems, for example, radiative transport theory:

$$u(x) = \int_0^{\infty} k(x-y)\,u(y)\,dy + f(x), \quad x > 0. \tag{20.2.3}$$

Here $k$ is assumed to be given, and integrable, on the whole line, but $f$ is given, and $u$ is sought, only on the half-line. Equation (20.2.3) can be converted to something closer to (20.2.1) by extending the (unknown) function $u$ to the whole line, and setting

$$u_+(x) = \begin{cases} u(x), & x \ge 0; \\ 0, & x < 0. \end{cases} \qquad u_-(x) = \begin{cases} 0, & x \ge 0; \\ u(x), & x < 0. \end{cases} \tag{20.2.4}$$

Let $f(x) = 0$, $x < 0$, i.e. $f = f_+$. Then (20.2.3) takes the form

$$u(x) = [k * u_+](x) + f_+(x). \tag{20.2.5}$$

This implicitly defines $u_-$ in terms of $u_+$:

$$u_- = [k * u_+]_-. \tag{20.2.6}$$

Note that $u_+ \in L^1$ implies also that $u_- \in L^1$.

For notational reasons, which will become clear in a moment, we choose a different normalization of the Fourier transform. This change will have no effect on any of the previous results; it merely amounts to replacing the previous $\widehat{f}(\xi)$ with $\widehat{f}(-\xi)$. Thus we set

$$\widehat{u}(\xi) \;=\; \int_{-\infty}^{\infty} u(x)\, e^{ix\xi}\, dx. \tag{20.2.7}$$

The renormalized inverse transform is

$$u(x) \;=\; \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ix\xi}\,\widehat{u}(\xi)\, d\xi. \tag{20.2.8}$$

The transformed version of (20.2.5) is

$$(1-\widehat{k})\,\widehat{u_+} \;=\; \widehat{f} - \widehat{u_-}. \tag{20.2.9}$$

The next step is to distinguish between terms like $\widehat{u_+}$ and $\widehat{u_-}$ as elements of $\mathscr{R}$.

**Proposition 20.2.1.** *Suppose that $u$ belongs to $L^1$ and $u_\pm$ are defined by (20.2.4). Then $\widehat{u_+}$ extends to a function $U^+$ that is holomorphic in the upper half plane $\mathbb{C}_+$ and continuous on the closure $\mathbb{C}_+ \cup \mathbb{R}$. Similarly $\widehat{u_-}$ extends to a function $U^-$ that is holomorphic in $\mathbb{C}_-$ and continuous on the closure $\mathbb{C}_- \cup \mathbb{R}$. Moreover, $U^\pm \to 0$ as $|z| \to \infty$ in the respective half-planes.*

*Proof:* In fact $|e^{ix\xi}| \;=\; \exp(-x\,\mathrm{Im}\,\xi)$ so $\mathrm{Im}\,\xi \geq 0$ implies that the function

$$U^+(\xi) \;\equiv\; \widehat{u_+}(\xi) \;=\; \int_0^{\infty} u(x)\, e^{ix\xi}\, dx$$

is well-defined and bounded by $||u_+||$, and similarly for $U^- = \widehat{u_-}$ in $C_-$. It is easily checked that the extensions are holomorphic. To prove the statements about continuity on the closure and vanishing at $\infty$ it is enough to consider the case of the indicator function of an interval; see Exercise 2. Then the result carries over to piecewise constant functions and, by approximation, to general functions in $L^1$. □

Let us note that there is an equivalent, but more direct, formulation of the splitting of $U$; see Exercise 1.

Proposition 20.2.1 leads to an equivalent version of (20.2.9):

$$(1-\widehat{k})\,U^+ \;=\; \widehat{f} - U^-. \tag{20.2.10}$$

The basic idea of Wiener and Hopf was to eliminate $U^-$ from (20.2.10) by *factoring* $A = 1-\widehat{k}$. We illustrate with a simple example. Consider a restricted case of Lalesco's equation

$$u(x) \;=\; \lambda \int_0^{\infty} e^{-|x-y|}\, u(y)\, dy + f(x), \quad x > 0, \quad \lambda < \frac{1}{2}. \tag{20.2.11}$$

Thus $k(x) = \lambda e^{-|x|}$ and

$$\widehat{k} = \lambda \int_{-\infty}^{0} e^{x+ix\xi}\, dx + \lambda \int_{0}^{\infty} e^{-x+ix\xi}\, dx$$
$$= \lambda \left\{ \frac{1}{1+i\xi} + \frac{1}{1-i\xi} \right\} = \frac{2\lambda}{1+\xi^2}. \tag{20.2.12}$$

Set $\mu = \sqrt{1-2\lambda} > 0$, so

$$1 - \widehat{k}(\xi) = \frac{\xi^2+\mu^2}{\xi^2+1} = \frac{(\xi+i\mu)(\xi-i\mu)}{(\xi+i)(\xi-i)}$$

and we have a factorization

$$A \equiv 1 - \widehat{k} = \frac{A^+}{A^-}, \qquad A^+ = \frac{\xi+i\mu}{\xi+i}, \quad A^- = \frac{\xi-i}{\xi-i\mu}.$$

Note that $A^{\pm}$ is holomorphic in $\mathbb{C}_{\pm}$. When $f \equiv 0$ in (20.2.11), equation (20.2.10) is equivalent to

$$A^+U^+ = -A^-U^-.$$

The left side of this equation extends holomorphically to $C_+$, while the right side extends holomorphically to $C_-$ and each side is continuous up to $\mathbb{R}$. Thus, taken together, they define an entire function $G$. (See the argument used in the proof of Theorem 1.6.1.) Each side has limit zero as $|\xi| \to \infty$. By Liouville's theorem $G = 0$. Thus $A^+U^+ = 0$; we have eliminated $U^-$ from the problem. Since $A^+$ has no real zeros, it follows that (20.2.11) has only the trivial solution $u = 0$. As we shall see, the argument can be extended to the inhomogeneous equation ($f \neq 0$), to show that (20.2.11) has a unique solution.

Wiener and Hopf generalized the procedure used in this example. Starting with (20.2.10), set $A = 1 - \widehat{k}$. Suppose, as in the example, that $A$ has no zeros on $\mathbb{R}$. Then a continuous determination of the logarithm $\log A(\xi)$ can be made along $\mathbb{R}$, normalized with

$$\lim_{\xi \to -\infty} \log A(\xi) = 0.$$

The limit as $\xi \to +\infty$ is $2m\pi i$ for some index $m \in \mathbb{Z}$. By definition, the integer $m$ is the *index* of $A$. Geometrically the index is $i$ times the change in the argument of $A(\xi)$ as $\xi$ goes from $-\infty$ to $+\infty$.

**Theorem 20.2.2.** *Suppose that $k$ and $f$ belong to $L^1$. Suppose that $A = 1 - \widehat{k}$ has no zeros on $\mathbb{R}$ and has index zero. Then (20.2.9) has a unique solution $u_+ \in L^1$.*

*Proof:* Since $A$ has no zeros on $\mathbb{R}$, and has index zero, the image $A(\mathbb{R})$ is a closed curve that passes through $z = 1$ and avoids the origin. Therefore the principal branch of the logarithm is holomorphic in a neighborhood of $A(\mathbb{R})$. By Theorem 20.1.3, $\log A$ belongs to $\mathscr{R}_0$. Let $L = L^+ + L^-$ be the corresponding splitting of $L = \log A$. Then the functions

$$A^+ = \exp L^+, \qquad A^- = \exp(-L^-) \tag{20.2.13}$$

belong to $\mathscr{R}$. Note that $A = A^+/A^-$. Moreover $A^\pm$ has a holomorphic extension to $\mathbb{C}_\pm$ and these extensions have no zeros on the closure of $\mathbb{C}_\pm$, respectively. Arguing exactly as in the example above, we see that (20.2.9) is equivalent to

$$A^+U^+ - H^+ \;=\; H^- - A^-U^-, \tag{20.2.14}$$

where $H = H^+ + H^-$ is the splitting of $H = A^-\widehat{f}$. As in the example, (20.2.14) determines an entire function that is identically zero. Thus $A^+U^+ = H^+$ and the solution $u_+$ of (20.2.9) is the inverse transform of

$$U^+ \;=\; \frac{H^+}{A^+} \;\in\; \mathscr{R}_0. \qquad \square$$

When the index of $A$ is not zero, it is necessary to compensate. Let

$$\Pi_m \;=\; \Pi_m(\xi) \;=\; \frac{(\xi+i)^m}{(\xi-i)^m}, \quad m = \pm 1, \pm 2, \pm 3, \dots .$$

As $\xi$ goes from $-\infty$ to $+\infty$, the argument of $\xi - i$ decreases by $\pi$, and the argument of $\xi + i$ increases by $\pi$. Thus $\Pi_m$ has index $m$. Note that $|\Pi_m(\xi)| = 1$, $\xi \in \mathbb{R}$. Note also that $\Pi_m$ belongs to $\mathscr{R}$; see Proposition 20.1.1.

**Theorem 20.2.3.** *Suppose that $k$ and $f$ belong to $L^1$. Suppose also that $A = 1 - \widehat{k}$ has no zeros on $\mathbb{R}$ and has index $m > 0$. Then (20.2.9) has an $m$-dimensional space of solutions $u \in L^1$.*

*Proof:* The index

$$\operatorname{ind}\left(\frac{A}{\Pi_m}\right) \;=\; \operatorname{ind} A - \operatorname{ind} \Pi_m \;=\; 0,$$

so the argument in Theorem 20.2.2 applies: $A/\Pi_m$ has a factorization $A^+/A^-$, where $A^\pm$ has the same properties as in the proof of Theorem 20.2.2. Then

$$A \;=\; \frac{A^+}{A^-}\frac{(\xi+i)^m}{(\xi-i)^m} \;=\; \frac{B^+}{B^-},$$

where

$$\begin{aligned} B^+ &= A^+(\xi+i)^m \;=\; \xi^m + O(|\xi|^{m-1}), \\ B^- &= A^-(\xi-i)^m \;=\; \xi^m + O(|\xi|^{m-1}) \end{aligned} \tag{20.2.15}$$

as $\xi \to \infty$. Note that $B^+$ has no zeros on $\mathbb{R}$. Let $H^\pm$ be defined as before: $A^-\widehat{f} = H^+ + H^-$. Then our equation takes the form

$$B^+U^+ - (\xi-i)^m H^+ \;=\; (\xi-i)^m H^- - B^-U^-.$$

Again the left side extends to the upper half plane and the right side extends to the lower half plane. It follows from (20.2.15) that the entire function thus defined is a

polynomial of degree less than $m$. Conversely, if $P$ is such a polynomial, the rational function $(\xi - i)^{-m}P/A^+\Pi_m$ belongs to $\mathscr{R}_0$. Therefore the inverse transform $u_+$ of

$$U^+ = \frac{(\xi-i)^m H^+ + P}{B^+} = \frac{H^+ + (\xi-i)^{-m}P}{A^+\Pi_m} \in \mathscr{R}_0$$

is a solution of (20.2.9). □

**Theorem 20.2.4.** *Suppose that $k$ and $f$ belong to $L^1$. Suppose also that $A = 1 - \widehat{k}$ has no zeros on $\mathbb{R}$ and has index $-m < 0$. Then (20.2.9) has a solution $u$ in $L^1$ if and only if $f$ satisfies a certain system of $m$ linearly independent constraints. If the solution exists, it is unique.*

*Proof:* In this case it is convenient to replace $\prod(\xi+i)^m(\xi-i)^{-m}$ by another product with index $m$, say

$$P_m(\xi) = \prod_{\nu=1}^{m} \frac{\xi + i\nu}{\xi - i\nu}.$$

Then $AP_m$ has index zero, so it has a factorization $A^+/A^-$. As in the previous proof, (20.2.9) becomes

$$\frac{A^+}{P_m}U^+ - H^+ = H^- - A^-U^-.$$

Both sides vanish at $\infty$, so the implicitly defined entire function is zero. The first term on the left has simple zeros at $\xi = i\nu$, $1 \le \nu \le m$, so there is a solution $U \in \mathscr{R}_0$ if and only if $H^+(i\nu) = 0$, $1 \le \nu \le m$. Since $H^+$ depends linearly on $f$, these can be taken to be the linear constraints referred to above. □

**Remark.** The linear constraints on $f$ in Theorem 20.2.4 can be made more explicit, in the form

$$\int_0^\infty f(x)\,w_\nu(x)\,dx = 0, \quad 1 \le \nu \le m, \tag{20.2.16}$$

where the $w_\nu$ are certain linearly independent bounded functions; see Exercises 4 and 7.

## 20.3 The case of real zeros of $1 - \widehat{k}$

In many physical problems the function $k$ is not only real but even. This implies that $\hat{k}$ is also real and even; Exercise 8. In particular $A = 1 - \hat{k}$ is real. If $A$ has no zeros, then the index is zero and Theorem 20.2.2 applies. Otherwise $A$ has an even number $2m$ of real zeros, counting multiplicity.

Suppose, somewhat more generally, that $k$ and $f$ belong to $L^1$, that $\widehat{k}$ is real-valued, and that $A = 1 - \hat{k}$ has zeros $a_1, a_2, \ldots a_{2m}$, counting multiplicity, in the sense that $A/Q$ is bounded and nowhere zero on $\mathbb{R}$, where

$$Q(\xi) = \prod_{j=1}^{2m}(\xi - a_j).$$

Let

$$B = \frac{A}{Q}(1+\xi^2)^m = A\,\frac{(\xi+i)^m(\xi-i)^m}{Q}.$$

Then $B$ is real on $\mathbb{R}$, with no zeros.

Under a mild smoothness assumption, the Hölder condition

$$|B(\xi)-B(\eta)| \le C|\xi-\eta|^\alpha, \qquad \text{some } 0<\alpha<1,$$

there is a factorization

$$B = \frac{B^+}{B^-}, \quad B^\pm \text{ holomorphic in } \mathbb{C}_\pm, \text{continuous on the closure,}$$

where $B_\pm$ has no zeros and has limit 1 as $\xi \to \infty$; see Exercise 5 of Chapter 18 and Exercise 1. Then

$$A = \frac{BQ}{(\xi+i)^m(\xi-i)^m} = \frac{B_+(\xi+i)^{-m}}{B_-(\xi-i)^m}\,Q = \frac{A^+}{A^-}\,Q,$$

where $A^\pm$ has no zeros in the closure of $\mathbb{C}^\pm$, and $A^\pm(z) = z^{\mp m}[1+O(|z|^{-1}]$ as $z\to\infty$ in $\mathbb{C}_\pm$.

The equation

$$AU^+ = \widehat{f}, \quad f \in L^1, \tag{20.3.1}$$

is the same as

$$A^+QU^+ = A^-\widehat{f} = H^+ + H^-, \qquad H^\pm = (A^-\widehat{f})^\pm.$$

Then

$$A^+QU^+ - H^+ = H^-$$

on $\mathbb{R}$. It follows that $A^+QU^+ - H^+$ extends as an entire function. This function is $o(z^m)$ at $\infty$, and thus is a polynomial $P$ of degree $< m$. Therefore our solution is, formally,

$$U^+ = \frac{H^+ + P}{A^+Q} = \frac{K}{Q}, \qquad K = \frac{H^+ + P}{A^+}. \tag{20.3.2}$$

Taking the inverse transform, we have, formally,

$$u(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ix\xi}\,\frac{K(\xi)}{Q(\xi)}\,d\xi, \qquad x>0. \tag{20.3.3}$$

The integrand in (20.3.3) has singularities at $\xi = a_j$, so the integral needs some clarification. For simplicity, let us suppose that the $a_j$ are distinct. As a first step, we use the fact that the integrand is holomorphic in $\mathbb{C}_+$ to modify the path of integration to avoid the singularities. Let $\Gamma$ be the union of

$$\mathbb{R} \setminus S_\delta, \qquad S_\delta = \bigcup_{j=1}^{2m} \{\xi : |\xi - a_j| \le \delta\}$$

and the semicircles $\Gamma_j = \{z : |z - a_j| = \delta, \operatorname{Im} z > 0\}$, where $\delta > 0$ is chosen small enough so that the $\Gamma_j$ are disjoint. The resulting integral is independent of $\delta$.

Near $a_j$, the integrand in (20.3.3) is approximately

$$\frac{K(a_j)}{Q(\xi)} = K(a_j)\,\frac{b_j}{\xi - a_j}, \qquad b_j = \prod_{k \ne j} (a_j - a_k)^{-1}.$$

Let us write

$$\frac{K(z)}{Q(z)} = \sum_{j=1}^{2m} \frac{b_j K(a_j)}{z - a_j} + L(z).$$

If $K$ satisfies the Hölder condition at the $a_j$, i.e. if for some $\beta > 0$, and for $j = 1, 2, \ldots, 2m$,

$$|K(\xi) - K(a_j)| \le C|\xi - a_j|^\beta, \qquad \text{if } |\xi - a_j| < \delta,$$

then the inverse transform $l$ of $L$ is well-defined.

It remains to determine the integral

$$\frac{1}{2\pi} \int_\Gamma e^{-ixz} \sum_{j=1}^{2m} \frac{b_j K(a_j)}{z - a_j}\, dz. \tag{20.3.4}$$

The integrand is holomorphic and decreasing at $\infty$ in the region below $\Gamma$, in such a way that the residue theorem applies. Because of the orientation of $\Gamma$, the integral (20.3.4) picks up $-i$ times the sum of the residues. Thus (20.3.2) is

$$u(x) = l(x) - i \sum_{j=1}^{2m} b_j K(a_j)\, e^{-ia_j x}. \tag{20.3.5}$$

## Exercises

1. Suppose that $U : \mathbb{R} \to \mathbb{C}$ is continuously differentiable and

$$\int_{-\infty}^{\infty} \frac{|U(\xi)|}{1 + |\xi|}\, d\xi < \infty.$$

Show that the splitting $U = U^+ + U^-$ as in Proposition 20.2.1 can be accomplished via the Cauchy transform:

$$U^\pm(\xi) = \pm \lim_{\varepsilon \to 0+} Cu(\xi \pm i\varepsilon) = \pm \lim_{\varepsilon \to 0+} \left\{ \frac{1}{2\pi i} \int_0^\infty \frac{U(t)\, dt}{t - (\xi \pm i\varepsilon)} \right\}.$$

2. Compute the Fourier transform of the indicator function of an interval:

$$f(x) \;=\; 1 \;\text{ if } a < x < b, \qquad f(x) \;=\; 0 \;\text{ otherwise.}$$

Note that $\widehat{f}(z)$ vanishes at $\infty$ in the half plane $\mathbb{C}_\pm$ if the interval $[a,b]$ is in the appropriate half-line $\mathbb{R}_\pm$.

3. (a) Let $f(x) = e^{-\mu|x|}$, $\mu > 0$.

(a) Show that $\widehat{f_+} = i/(\xi + i\mu)$.

(b) Show that $\widehat{f_-} \;=\; -i/(\xi - i\mu)$.

(c) Suppose $\operatorname{Im}\alpha > 0$. Find the inverse transform of $1/(\xi+\alpha)$.

(d) Suppose $\operatorname{Im}\alpha < 0$. Find the inverse transform of $1/(\xi-\alpha)$.

(e) Show that the inverse transform of $1/(1+\xi^2)$ is $\pi e^{-|x|}$.

4. (a) Suppose that $f$ and $g$ belong to $L^1$. Suppose that $f$ is bounded. Show that the convolution $f * g$ (which belongs to $L^1$) is also bounded.

(b) Deduce from Exercise 3 that if $\alpha$ is not real, then the inverse transform of $1/(\xi-\alpha)$ is a bounded $L^1$ function.

(c) Suppose that $V$ belongs to $\mathscr{R}_0$, while $R$ is a rational function that vanishes at $\infty$ and has no poles on $\mathbb{R}$. Prove that the inverse transform of the product $VR$ is a bounded $L^1$ function.

5. Suppose that $g$ is a Schwartz function and that $G = \widehat{g}$. Prove that

$$\sup|G(\xi)| \;\leq\; ||g||_1 \;\leq\; \sup|G(\xi)| + \sup|G'(\xi)|.$$

Use this to give a direct proof of Theorem 20.1.2. Suppose that $F = \widehat{f}$ belongs to $\mathscr{R}$. Choose a Schwartz function $G$ such that $||G-F|| \leq \frac{1}{2}$. Prove that the series

$$\sum_{n=0}^{\infty} \frac{(G-F)^n}{G^{n+1}}$$

converges in norm in $\mathscr{R}$, with limit $1/f$. (This is an adaptation to the Fourier integral case of a proof by Newman [106] of Wiener's theorem in the case of Fourier series; see Exercises 14 and 15 of Chapter 4.)

6. Under the assumptions of Theorem 20.2.3, show that each solution $w$ in $L^1$ of the homogeneous equation

$$w(x) \;=\; \int_0^\infty k(x-y)\,w(y)\,dy, \quad x > 0,$$

is a bounded function.

7. Under the assumptions, and notation, of Theorem 20.2.4, let $\widetilde{k}(x) = k(-x)$.

(a) Show that the transform of $\widetilde{k}$ is $1-\widetilde{A}$, where $\widetilde{A}(\xi) = A(-\xi)$.

(b) Show that $\widetilde{A}$ has index $m$.

(c) Let $w$ be a solution of the homogeneous equation

$$w(x) \;=\; \int_0^\infty \widetilde{k}(x-y)\,w(y)\,dy \;=\; \int_0^\infty k(y-x)\,w(y)\,dy, \quad x>0.$$

Show that if (20.2.9) has a solution $u \in L^1$, then

$$\int_0^\infty f(x)\,w(x)\,dx \;=\; 0.$$

(d) Show that there are $m$ linearly dependent solutions of the homogeneous equation.

8. Suppose that $f \in L^1$ is even and real-valued. Show that $\hat{f}$ is even and real-valued.
9. (a) For what values of the parameter $\lambda$ does the Lalesco problem

$$u(x) \;=\; \lambda \int_0^\infty e^{-|x-y|} u(y)\,dy, \qquad x>0.$$

have a non-trivial solution?

(b) Find the non-trivial solutions.

10. The Milne equation comes from the theory of radiative equilibrium. (For a derivation, see Milne [98] or Titchmarsh [136].) The equation is

$$u(x) \;=\; \int_0^\infty k(x-y)\,u(y)\,dy, \qquad k(x) \;=\; \frac{1}{2}\int_{|x|}^\infty \frac{e^{-t}}{t}\,dt. \tag{20.3.6}$$

(a) Show that $\widehat{k}(\xi) \;=\; \tan^{-1}(\xi)/\xi$. (It may help to note that

$$\frac{\sin(\xi t)}{t} \;=\; \int_0^\xi \cos(st)\,ds.)$$

(b) Show that $A = 1 - \widehat{k}$ has a double zero at $\xi = 0$ and no other real zeros.

(c) Discuss the solutions of (20.3.6).

## Remarks and further reading

Wiener and Hopf [144] assumed for $k$ the decay condition

$$|k(x)| \;=\; O(e^{-\varepsilon|x|}) \quad \text{as} \;\; |x| \to \infty$$

for some $\varepsilon > 0$. This implies that $\widehat{k}_\pm$ are both holomorphic in the strip $\{z : |\mathrm{Im}\, z| < \varepsilon\}$. This is the route taken in many treatments, such as Paley and Wiener [115], Chapter IV; Titchmarsh [136], §11.17; and Noble [109]. We have followed instead the approach of Krein [80] and Gohberg and Krein [49], who only assume that $k$ belongs to $L^1$. This latter approach is conceptually simpler, at the price of invok-

ing more powerful (but important) machinery, namely, Theorem 20.1.2 and Theorem 20.1.3.

Widom [141] contains a discussion of the original technique and subsequent developments involving operator theory. For much more on the subject from the point of view of applications, see the articles in Lawrie and Abrahams [84].

# Chapter 21
# Tauberian theorems



This chapter is largely an exercise in real analysis, but there are important connections to complex topics and methods. Moreover the starting point of the theory is a theorem of Abel concerning the boundary behavior of a function holomorphic in the unit disk:

**Theorem 21.0.1.** *If $\sum_{n=0}^{\infty} a_n z^n$ is holomorphic in the unit disk, let*

$$f(x) = \sum_{n=1}^{\infty} a_n x^n, \quad -1 < x < 1. \tag{21.0.1}$$

*If the series $\sum_{n=0}^{\infty} a_n$ converges, then*

$$\lim_{x \to 1-} f(x) = \sum_{n=1}^{\infty} a_n. \tag{21.0.2}$$

The converse is false, in general. If $a_n = (-1)^n$ then $f(z) = 1/(1+z)$ has limit $1/2$ at $z = 1$, but the series does not converge. A necessary condition for convergence is that $a_n \to 0$. In 1897 Tauber showed that a sufficient condition for the existence of $\lim_{x\to 1-} f(x)$ to imply convergence is that $a_n = o(1/n)$. In 1906 Fatou [46] showed that the necessary condition $a_n \to 0$ is also sufficient if $f$ has a holomorphic extension to a neighborhood of $z = 1$.

Some years later Hardy and Littlewood began a series of investigations inspired by Tauber's theorem. They used the term "Tauberian" to refer to results of this type, where an additional condition allows one to prove the converse of a relatively easy "abelian" theorem like that of Abel. In particular, Littlewood sharpened Tauber's condition to $a_n = O(1/n)$.

Generally speaking, an abelian theorem has the following character: One type of convergence of a sequence or a function implies a weaker type. A tauberian theorem gives (preferably optimal) conditions under which the weaker convergence implies the stronger. The simplest example is that of a numerical sequence $\{s_n\}_{n=1}^{\infty}$. It is



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_21

easy to show that if

$$\lim_{n\to\infty} s_n \;=\; s \tag{21.0.3}$$

then the averages $\{A_n\}$ of the first $n$ elements also converge to $s$:

$$\lim_{n\to\infty} A_n \;=\; s, \qquad A_n \;=\; \frac{s_1+s_2+\cdots+s_n}{n}. \tag{21.0.4}$$

The example $s_n = (-1)^n$ shows that (21.0.4) does not imply (21.0.3). One necessary condition for (21.0.3) is that the successive differences $s_n - s_{n-1}$ converge to zero. A theorem of Hardy [59] gives a sharp condition: if $s_{n+1} - s_n = O(n^{-1})$ and (21.0.4) holds, then so does (21.0.3). (This can fail if $s_{n+1} - s_n$ decays more slowly, e.g. if $s_{n+1} - s_n \;=\; O(n^{-1+\varepsilon})$.) Hardy's condition can be thought of as a limitation on the oscillation of the sequence. A second type of tauberian condition rules out oscillation entirely. For example, if the sequence $\{s_n\}$ is non-decreasing, then (21.0.4) implies (21.0.3); see Exercise 2.

In this chapter we prove Hardy's result in a continuous version, and then turn to theorems of Abel, Littlewood, and Hardy–Littlewood, and to Karamata's proof of the Hardy–Littlewood result.

Section 21.4 on the Wiener tauberian theorem depends on Sections 18.3, 18.4. The present chapter concludes with a tauberian theorem of Malliavin which takes advantage of complex information to obtain an error estimate.

## 21.1 Hardy's theorem

The proof of the theorem of Hardy, mentioned in the introduction above, may be clearer in a continuous version. Suppose that $f : [0,\infty) \to \mathbb{C}$ is continuous and piecewise differentiable. Let

$$g(x) = \int_0^x f(t)\,dt.$$

The analogue of (21.0.4) is

$$\lim_{x\to\infty} \frac{g(x)}{x} \;=\; A, \tag{21.1.1}$$

and the analogue of the condition $s_{n+1} - s_n = O(n^{-1})$ is

$$|f(x+h) - f(x)| \;\le\; \frac{Kh}{x}, \quad \text{if } \; h \ge x \ge x_0. \tag{21.1.2}$$

We show here that, under these conditions,

$$\lim_{x\to\infty} f(x) \;=\; A. \tag{21.1.3}$$

Hardy's theorem, as stated in the introduction, can be derived from this by taking $f(x) = s_n$ for $n-1 < x \le n$, or by carrying out the discrete analogues of the following calculations, i.e. with summations in place of integrations.

Given $\delta > 0$, for sufficiently large $x$ we have $|g(x) - Ax| \le \delta x$. Therefore for $h > 0$,

$$\begin{aligned}\left|\frac{g(x+h)-g(x)}{h} - A\right| &= \left|\frac{g(x+h)-A(x+h)}{h} - \frac{g(x)-Ax}{h}\right| \\ &\le \frac{2x\delta}{h}.\end{aligned} \tag{21.1.4}$$

Thus, since $\delta > 0$ can be chosen at our discretion, we see that

$$\frac{x}{h} \le \text{constant} \quad \Rightarrow \quad \lim_{x\to\infty} \frac{g(x+h)-g(x)}{h} = A. \tag{21.1.5}$$

On the other hand, using (21.1.2) we have

$$\begin{aligned}\left|\frac{g(x+h)-g(x)}{h} - f(x)\right| &= \left|\frac{1}{h}\int_0^h f(x+t)\,dt - f(x)\right| \\ &= \left|\frac{1}{h}\int_0^h [f(x+t)-f(x)]\,dt\right| \\ &\le \left|\frac{1}{h}\int_0^h \frac{Kt}{x}\,dt\right| \\ &= \frac{1}{h}\,\frac{Kh^2}{2x} = \frac{Kh}{2x}.\end{aligned} \tag{21.1.6}$$

Given $\varepsilon > 0$, the difference (21.1.6) is $\le \varepsilon$ if $x/h = K/2\varepsilon$. Together with (21.1.5), this implies that $f(x)$ has limit $A$. $\square$

Condition (21.1.2) is implied by the condition $f'(x) = O(x^{-1})$. The example

$$g(x) = \sin(x^\varepsilon)\,x^{1-\varepsilon}, \quad \varepsilon > 0, \tag{21.1.7}$$

shows that (21.0.3) can fail if we only require $|f'(x)| = O(x^{-1+\varepsilon})$ for some $\varepsilon > 0$; see Exercise 3.

## 21.2 Abel, Tauber, Littlewood, and Hardy–Littlewood

The proof of Abel's theorem, that if $f(z) = \sum_{n=0}^\infty a_n z^n$, then

$$\sum_{n=0}^\infty a_n = A \quad \Rightarrow \quad \lim_{z\to 1} f(z) = A,$$

is straightforward. Let $\{s_n\}$ be the sequence of partial sums of the series, and set $s_{-1} = 0$. Then for $|x| < 1$,

$$f(x) = \sum_{n=0}^{\infty} (s_n - s_{n-1}) x^n = (1-x) \sum_{n=0}^{\infty} s_n x^n.$$

Since also $(1-x)\sum_{n=0}^{\infty} x^n = 1$ if $|x| < 1$, we have, for $0 < x < 1$,

$$\begin{aligned} |f(x) - A| &= \left| (1-x) \sum_{n=0}^{\infty} (s_n - A) x^n \right| \\ &\leq (1-x) \left| \sum_{n=0}^{N} (s_n - A) \right| + \sup_{n>N} |s_n - A|. \end{aligned}$$

For large $N$ the second term on the right is small, and for fixed $N$ the first term approaches zero as $x \to 1$. □

As stated in the introduction, the first converse result is due to Tauber [133].

**Theorem 21.2.1.** (Tauber) *If $a_n = o(1/n)$, then*

$$\lim_{x \to 1-} \sum_{n=0}^{\infty} a_n x^n = A \quad \Rightarrow \quad \sum_{n=0}^{\infty} a_n = A.$$

*Proof:* Again let $\{s_n\}$ be the partial sums and let $f$ be defined by (21.0.1). Then for $|x| < 1$,

$$s_n - f(x) = \sum_{k=1}^{n} a_k (1 - x^k) - \sum_{k=n+1}^{\infty} a_k x^k.$$

Since $|x| < 1$ implies $1 - x^k = (1-x)(1 + x + \cdots + x^{k-1}) < k(1-x)$, we have

$$|s_n - f(x)| < (1-x) \sum_{k=1}^{n} k|a_k| + \sum_{k=n+1}^{\infty} |a_k| x^k.$$

By assumption, given $\varepsilon > 0$ there is $N$ such that $n > N$ implies $n|a_n| < \varepsilon$. For such $n$,

$$\sum_{k=n+1}^{\infty} |a_k| x^k < \varepsilon \sum_{k=n+1}^{\infty} \frac{x^k}{k} < \frac{\varepsilon}{n} \sum_{k=0}^{\infty} x^k = \frac{\varepsilon}{n(1-x)}.$$

Let $x_n = 1 - 1/n$. The previous estimates show that

$$\begin{aligned} |s_n - f(x_n)| &< \frac{1}{n}\sum_{k=1}^{n} k|a_k| + \varepsilon \\ &\leq \frac{1}{n}\sum_{k=1}^{N} k|a_k| + \frac{1}{n}\left\{\sum_{k=N+1}^{n}\right\}\varepsilon + \varepsilon \\ &= \frac{1}{n}\sum_{k=1}^{N} k|a_k| + \left(\frac{n-N}{n}\right)\varepsilon + \varepsilon. \end{aligned}$$

Taking $n \to \infty$, we have $f(x_n) \to A$ and the right side has limit $2\varepsilon$, so $s_n \to A$. □

The proof of Fatou's result, mentioned in the introduction, depends on the localization principle for Fourier series; see [149], IX.4.3, IX.5.7.

Hardy conjectured that the sharp condition for the converse of Abel's theorem is $a_n = O(1/n)$. Littlewood [93] proved this in 1911.

**Theorem 21.2.2.** (Littlewood) *If* $a_n = O(1/n)$, *then*

$$\lim_{x\to 1-}\sum_{n=0}^{\infty} a_n x^n = A \quad \Rightarrow \quad \sum_{n=0}^{\infty} a_n = A.$$

Littlewood's proof was considerably more difficult than Tauber's proof of the weaker result. In 1914 Hardy and Littlewood [60] proved a different type of result, using positivity (non-oscillation).

**Theorem 21.2.3.** (Hardy–Littlewood) *If each* $a_n \geq 0$, *then*

$$\lim_{x\to 1-}(1-x)\sum_{n=1}^{\infty} a_n x^n = A \quad \Rightarrow \quad \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N} a_n = A.$$

Littlewood's theorem could be deduced from this result, making use of Tauber's theorem, Theorem 21.2.1, in the process.

The proof of Theorem 21.2.3 was greatly simplified, and the result itself considerably generalized, by Karamata [75].

## 21.3 Karamata's tauberian theorem

Karamata's transformation of the original problem is instructive: one tauberian theorem may be essentially the same as another, in a different guise.

Let us rewrite the series in (21.0.1) by setting

$$x = e^{-\varepsilon}, \qquad s(t) = \sum_{n\leq t} a_n.$$

Then the series takes the form of a Stieltjes integral (Section 1.8):

$$\sum_{n=0}^{\infty} a_n x^n = \int_0^\infty e^{-\varepsilon t}\, ds(t) = \frac{1}{\varepsilon}\int_0^\infty e^{-\varepsilon t} s(t)\, dt, \qquad (21.3.1)$$

and Theorem 21.2.3 is the statement that if each $a_n \geq 0$, then

$$\lim_{\varepsilon\to 0+}\int_0^\infty e^{-\varepsilon t} s(t)\, dt = A \quad \Rightarrow \quad \lim_{N\to\infty}\frac{s(N)}{N} = A.$$

This is a special case of Theorem 21.3.1. We need here a few simple facts about the gamma function, defined by

$$\Gamma(a) = \int_0^\infty e^{-t} t^{a-1}\, dt, \qquad a > 0.$$

It follows easily from an integration by parts and a change of variables that for $a > 0$, $m > 0$,

$$a\Gamma(a) = \Gamma(a+1), \qquad \int_0^\infty e^{-mt} t^{a-1}\, dt = \frac{\Gamma(a)}{m^a}.$$

Let us start with an example. If $\alpha(t) = t^a$, $a > 0$, then

$$\begin{aligned}\varepsilon^a \int_0^\infty e^{-\varepsilon t}\, d\alpha(t) &= a\,\varepsilon^a \int_0^\infty e^{-\varepsilon t} t^{a-1}\, dt \\ &= a\int_0^\infty e^{-s} s^{a-1}\, ds = a\Gamma(a) = \Gamma(a+1). \qquad (21.3.2)\end{aligned}$$

It is not difficult to extend this:

$$\lim_{t\to\infty}\frac{\alpha(t)}{t^a} = 1 \quad \Rightarrow \quad \lim_{\varepsilon\to 0+} \varepsilon^a \int_0^\infty e^{-\varepsilon t}\, d\alpha(t) = \Gamma(a+1); \qquad (21.3.3)$$

Exercise 4. Karamata's theorem is a partial converse that applies if $\alpha$ is non-decreasing.

**Theorem 21.3.1.** (Karamata) *Suppose that* $\alpha : [0,\infty) \to \mathbb{R}$ *is non-decreasing, and for some* $a > 0$,

$$\lim_{\varepsilon\to 0+}\left\{\varepsilon^a \int_0^\infty e^{-\varepsilon t}\, d\alpha(t)\right\} = 1. \qquad (21.3.4)$$

*Then*

$$\alpha(t) \sim \frac{t^a}{\Gamma(a+1)} \quad as\ \ t \to \infty. \qquad (21.3.5)$$

*Proof:* For each positive integer $m$,

$$\begin{aligned}\lim_{\varepsilon\to 0+}\left\{\varepsilon^a \int_0^\infty (e^{-\varepsilon t})^m\, d\alpha(t)\right\} &= \lim_{\varepsilon\to 0+}\left\{\frac{(m\varepsilon)^a}{m^a}\int_0^\infty e^{-m\varepsilon t}\, d\alpha(t)\right\} \\ &= \frac{1}{m^a} = \frac{1}{\Gamma(a)}\int_0^\infty e^{-mt} t^{a-1}\, dt. \qquad (21.3.6)\end{aligned}$$

Therefore for each polynomial $P$ such that $P(0) = 0$,

$$\lim_{\varepsilon \to 0} \left\{ \varepsilon^a \int_0^\infty P(e^{-\varepsilon t})\, d\alpha(t) \right\} = \frac{1}{\Gamma(a)} \int_0^\infty P(e^{-t})\, t^{a-1}\, dt. \tag{21.3.7}$$

It follows from Weierstrass's polynomial approximation theorem, Theorem 4.4.2, that any continuous function $f : [0,1] \to \mathbb{R}$ such that $f(0) = 0$ can be approximated uniformly by polynomials $P$ such that $P(0) = 0$. Therefore the identity (21.3.7) carries over, with $f$ in place of $P$. Choose a sequence of such continuous functions $\{f_n\}$ with the property that the $f_n$ are pointwise non-decreasing, and

$$\lim_{n \to \infty} f_n(s) = \begin{cases} 1 & \text{if } e^{-1} \le s < 1, \\ 0 & \text{if } 0 < s \le e^{-1}. \end{cases} \tag{21.3.8}$$

It follows from the assumption that $\alpha$ is non-decreasing that the integrals

$$\int_0^\infty f_n(e^{-\varepsilon t})\, d\alpha(t)$$

increase to the limit

$$\int_0^{1/\varepsilon} d\alpha(t) = \alpha(1/\varepsilon) - \alpha(0).$$

Therefore, from (21.3.7) and (21.3.8), we have

$$\begin{aligned} \lim_{\varepsilon \to 0+} \left\{ \varepsilon^a \int_0^{1/\varepsilon} d\alpha(t) \right\} &= \frac{1}{\Gamma(a)} \int_0^1 t^{a-1}\, dt \\ &= \frac{1}{a\Gamma(a)} = \frac{1}{\Gamma(a+1)}, \end{aligned}$$

which is (21.3.5). □

There are tauberian theorems for *Dirichlet series*

$$f(s) = \sum_{n=1}^\infty \frac{a_n}{n^s} = \sum_{n=1}^\infty \frac{a_n/n}{n^{s-1}}. \tag{21.3.9}$$

These can be put into the Karamata form (21.3.1) by setting

$$\alpha(t) = \sum_{n \le e^t} \frac{a_n}{n},$$

so that

$$f(s) = \int_0^\infty e^{-(s-1)t}\, d\alpha(t), \tag{21.3.10}$$

and the limit $s \to 1$ corresponds to the limit $\varepsilon \to 0+$ above.

The most famous result for Dirichlet series is a theorem of Ikehara [69] that was used to provide a relatively simple proof of the prime number theorem. Note that in this case one relies heavily on the behavior of $f$ as a complex function.

**Theorem 21.3.2.** (Ikehara) *Suppose that $f$ has the form (21.3.9) with $a_n \geq 0$. Suppose also that $f$ is holomorphic for* $\mathrm{Re}\, s > 1$ *and that the function $(s-1)f(s) - A$ has a continuous extension to the closed half plane $\{s : \mathrm{Re}\, s \geq 1\}$. Then*

$$\lim_{n\to\infty} \left\{ \frac{1}{n} \sum_{k=1}^{n} a_k \right\} = A.$$

This is closely related to the case $a = 1$ of Karamata's theorem. The stronger hypotheses give a stronger result; see Exercise 6.

Let us pass to the next section via one more transformation. Suppose that $\alpha$ has a bounded derivative $H$, so that

$$\varepsilon \int_0^\infty e^{-\varepsilon t}\, d\alpha(t) = \int_0^\infty F(\varepsilon t) H(t) \frac{dt}{t}. \tag{21.3.11}$$

Set

$$\varepsilon = e^{-x}, \quad t = e^y, \quad f(x) = F(e^{-x}), \quad h(x) = H(e^x). \tag{21.3.12}$$

Then (21.3.11) becomes

$$\int_{-\infty}^\infty f(x-y)\, h(y)\, dy. \tag{21.3.13}$$

Note that asymptotic behavior of (21.3.11) as $\varepsilon \to 0+$ corresponds to asymptotic behavior of (21.3.13) as $x \to \infty$. Note also that $f(x) = e^{-x} \exp(-e^{-x})$.

## 21.4 Wiener's tauberian theorem

This section depends on Sections 18.3 and 18.4. Let us recall some material from those sections. The Fourier transform

$$\widehat{f}(\xi) = \int_{-\infty}^\infty f(x)\, e^{-ix\xi}\, dx.$$

maps $L^1(\mathbb{R})$ into $C_0(\mathbb{R})$, the space of continuous functions with limit zero as $|x| \to \infty$.

We denote by $\mathscr{R}_0 \subset C_0(\mathbb{R})$ the image of $L^1$ under the Fourier transform. Multiplication in $\mathscr{R}_0$ corresponds to convolution in $L^1$:

$$h(x) = \int_{-\infty}^\infty f(x-y)\, g(y)\, dy \quad \Rightarrow \quad \widehat{h} = \widehat{f}\,\widehat{g}.$$

The norm in $\mathscr{R}_0$ is taken to be the $L^1$ norm of the inverse transform:

$$||\widehat{f}|| = \int_{-\infty}^\infty |f(x)|\, dx.$$

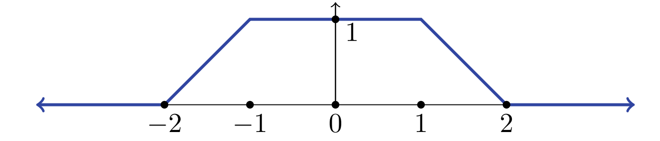



**Fig. 21.1** The function $U(\xi)$

Let $\mathscr{R}_c$ denote the subspace of $\mathscr{R}_0$ consisting of functions with compact support, i.e. functions $\widehat{f}$ such that $\widehat{f}(\xi) = 0$ for sufficiently large $|\xi|$.

The following piecewise linear function will play several roles here:

$$U(\xi) = \begin{cases} 0 & \text{if } \xi \le -2; \\ 2+\xi & \text{if } -2 < \xi < -1; \\ 1 & \text{if } -1 \le \xi \le 1 \\ 2-\xi & \text{if } 1 < \xi < 2; \\ 0 & \text{if } \xi \ge 2. \end{cases} \tag{21.4.1}$$

See Figure 21.1.

**Lemma 21.4.1.** *The function $U$ belongs to $\mathscr{R}_c$.*

*Proof:* The inverse transform

$$u(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} U(\xi)e^{ix\xi}\,d\xi = \frac{\cos x - \cos 2x}{\pi x^2} \tag{21.4.2}$$

is integrable on the line, so $U$ belongs to $\mathscr{R}$. By definition, $U$ has compact support.
□

Note that for $\varepsilon > 0$ and any integers $M < N$, the sum

$$\sum_{k=M}^{N} U\left(\frac{x-3\varepsilon k}{\varepsilon}\right) \tag{21.4.3}$$

is identically 1 on the interval $[a,b]$ and vanishes outside the interval $[a-\varepsilon, b+\varepsilon]$, where $a = (3M-1)\varepsilon$ and $b = (3N+1)\varepsilon$.

**Lemma 21.4.2.** *For each $f \in L^1$,*

$$\lim_{\varepsilon\to 0}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\big|[\varepsilon\, u(\varepsilon(x-y)) - \varepsilon\, u(\varepsilon x)]\, f(y)\big|\,dy\,dx = 0. \tag{21.4.4}$$

*Proof:* Letting $t = \varepsilon x$ and reversing the order of integration, the integral is

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \big|f(y)[u(t-\varepsilon y) - u(t)]\big|\,dt\,dy$$
$$= \iint_{|y|>N} + \iint_{|y|<N,|t|>2N} + \iint_{|y|<N,|t|<2N} = I_1 + I_2 + I_3.$$

Assume that $0 < \varepsilon \le 1$. Then $|y| < N$, $|t| > 2N$ implies that $|t - \varepsilon y| > |N|$. Therefore

$$|I_1| + |I_2| \le 2||u|| \int_{|y|>N} |f(y)|\,dy + ||f|| \cdot 2\int_{|t|>N} |u(t)|\,dt,$$

and both terms on the right have limit zero as $N \to \infty$. But also

$$|I_3| \le ||f|| \int_{|y|<N,|t|<2N} |u(t-\varepsilon y) - u(t)|\,dt,$$

and for each fixed $N$ the integrand converges uniformly to zero. □

**Lemma 21.4.3.** *The subring $\mathscr{R}_c$ is dense in $\mathscr{R}_0$.*

*Proof:* The inverse transform of the convolution $U * U$ is $2\pi u^2$, which is non-negative and integrable. In particular, the family of functions

$$G_\varepsilon(x) = \frac{1}{c\,\varepsilon}\,u\left(\frac{x}{\varepsilon}\right)^2, \qquad c = \int_{-\infty}^{\infty} u(x)^2\,dx$$

is easily seen to be an approximate identity. Therefore given any $f \in L^1$, the functions $G_\varepsilon * f$ converge to $f$ with respect to the $L^1$ metric. On the other hand, the Fourier transform of $G_\varepsilon * f$ is the product of $\widehat{f}$ with

$$\frac{1}{c}\,U * U(\varepsilon\xi),$$

which has compact support. Therefore $\widehat{f}$ is in the closure of $\mathscr{R}_c$. □

Finally we come to the tauberian theorem of Wiener [142].

**Theorem 21.4.4.** *If $F$ belongs to $\mathscr{R}_0$ and has no zeros, then the ideal*

$$F\mathscr{R}_0 = \{FG : G \in \mathscr{R}_0\} \tag{21.4.5}$$

*is dense in $\mathscr{R}_0$.*

*Proof:* Consider the quotient

$$\frac{U(\xi/\varepsilon)}{F(\xi)} = \frac{U(\xi/\varepsilon)}{F(0) + U(\xi/2\varepsilon)[F(\xi) - F(0)]}. \tag{21.4.6}$$

This identity holds because $U(\xi/\varepsilon) \neq 0$ implies that $U(\xi/2\varepsilon) = 1$. Now $U(\xi/2\varepsilon)$ is the Fourier transform of

$$u_\varepsilon(x) \equiv 2\varepsilon\, u(2\varepsilon x); \tag{21.4.7}$$

Exercise 7. Therefore $U(\xi/2\varepsilon)[F(\xi) - F(0)]$ is the Fourier transform of

$$u_\varepsilon * f(x) - u_\varepsilon(x)F(0) = \int_{-\infty}^{\infty} u_\varepsilon(x-y) f(y)\, dy - u_\varepsilon(x) \int_{-\infty}^{\infty} f(y)\, dy. \tag{21.4.8}$$

By Lemma 21.4.2 (with $2\varepsilon$ in place of $\varepsilon$), the $L^1$ norm of the function (21.4.8) is $\leq \frac{1}{2}|F(0)|$ for sufficiently small $\varepsilon$. Thus for such $\varepsilon$,

$$\left|\left| \frac{U(\xi/2\varepsilon)[F(\xi) - F(0)]}{F(0)} \right|\right| \leq \frac{1}{2}.$$

The identity (21.4.6) can be written

$$\frac{U(\xi/\varepsilon)}{F(\xi)} = \frac{U(\xi/\varepsilon)}{F(0)} \sum_{n=0}^{\infty} \left[ \frac{U(\xi/2\varepsilon)[F(\xi) - F(0)]}{F(0)} \right]^n,$$

and we have shown that for small $\varepsilon$, the series converges in norm. Therefore $V(\xi/\varepsilon) = U(\xi/\varepsilon)/F(\xi)$ belongs to $\mathscr{R}_0$ and

$$U(\xi/\varepsilon) = F(\xi)V(\xi/\varepsilon).$$

The same argument applies to translates, so for each $\xi_0$ there is a choice of $\varepsilon > 0$ and a function $W_{\xi_0} \in \mathscr{R}_0$ such that

$$U(\xi_0 + \xi/\varepsilon) = F(\xi) W_{\xi_0}(\xi/\varepsilon).$$

Suppose now that $G$ belongs to $\mathscr{R}_c$. Since $|F(\xi)|$ is bounded away from zero, the previous argument shows that for some $\varepsilon > 0$ we can find a sum $S$ of the form (21.4.3) that is identically 1 on the support of $G$, such that $S/F$ belongs to $\mathscr{R}_0$. Therefore

$$G = SG = F \cdot \frac{S}{F} G.$$

Thus the ideal $F\mathscr{R}_0$ contains $\mathscr{R}_c$, which is dense in $\mathscr{R}_0$ by Lemma 21.4.3. □

The non-vanishing condition is necessary. In fact suppose that $F \in \mathscr{R}_0$ vanishes at $\xi_0$. We may find $H \in \mathscr{R}_0$ such that $H(\xi_0) = 1$; for example choose a positive function $h_0$ with integral $\int_{-\infty}^{\infty} h_0(x)\, dx = 1$, let $h(x) = e^{ix\xi_0} h_0(x)$, and $H = \widehat{h}$. Then for each $G \in \mathscr{R}_0$,

$$||FG - H|| \geq |F(\xi_0)G(\xi_0) - H(\xi_0)| = 1.$$

It is not obvious why Theorem 21.4.4 is called a tauberian theorem. A second formulation makes it clearer, by showing that the non-vanishing of $F$ implies that certain averaging methods can be compared. We need one simple lemma.

**Lemma 21.4.5.** *Suppose that $g_1$ belongs to $L^1$ and $h:\mathbb{R}\to\mathbb{C}$ is bounded. If $h(x)$ has a limit as $x\to\infty$, then*

$$\lim_{x\to\infty}\int_{-\infty}^{\infty} g_1(x-y)\,h(y)\,dy \;=\; \lim_{x\to\infty} h(x)\int_{-\infty}^{\infty} g_1(y)\,dy. \tag{21.4.9}$$

*Proof:* Subtracting a constant, we may assume that $h(x)\to 0$. If $x>2N$ and $y<N$, then $x-y>N$ and

$$\begin{aligned}\int |g_1(x-y)\,h(y)|\,dy &= \int_{y<N}|g_1(x-y)\,h(y)|\,dy+\int_{y>N}|g_1(x-y)\,h(y)|\,dy\\ &\le \int_{t>N}|g_1(t)|\,dt\cdot\sup_y|h(y)|+\int_{-\infty}^{\infty}|g_1(t)|\,dt\cdot\sup_{y>N}|h(y)|.\end{aligned}$$

Both terms converge to zero as $N\to\infty$. □

**Theorem 21.4.6.** (Wiener tauberian theorem, version 2). *Suppose that $f$ and $g$ belong to $L^1$ and suppose that the Fourier transform $F$ of $f$ has no zeros. If $k$ is a bounded function such that*

$$\lim_{x\to\infty}\int_{-\infty}^{\infty} f(x-y)\,k(y)\,dy \;=\; A\int_{-\infty}^{\infty} f(y)\,dy, \tag{21.4.10}$$

*then also*

$$\lim_{x\to\infty}\int_{-\infty}^{\infty} g(x-y)\,k(y)\,dy \;=\; A\int_{-\infty}^{\infty} g(y)\,dy. \tag{21.4.11}$$

*Proof:* It is easily checked that the $L^1$ limit of a sequence of functions $g_n$ that satisfy (21.4.11) also satisfies (21.4.11). Therefore it is enough to show that the set of such functions is dense in $L^1$. Suppose the Fourier transform of $g$ belongs to $\mathscr{R}_c$. By the proof of Theorem 21.4.4 we know that there is a function $G_1$ in $\mathscr{R}_0$ such that $G=G_1F$. Let $g_1$ be the inverse Fourier transform of $G_1$. Then $g=g_1*f$, so Lemma 21.4.5 applies to $g_1$ and

$$h(x) \;=\; \int_{-\infty}^{\infty} f(x-y)\,k(y)\,dy$$

gives (21.4.11). □

As an illustration, suppose that the Fourier transform of $f$ has no zeros and $k$ is a bounded function such that (21.4.10) is true. Let $g$ be the function

$$g(x) \;=\; 1 \quad\text{if } -1<x<0, \qquad g(x) \;=\; 0, \quad\text{otherwise.}$$

Then

$$\int_{-\infty}^{\infty} g(x-y)\,k(y)\,dy \;=\; \int_{x}^{x+1} k(y)\,dy \;\sim\; A. \tag{21.4.12}$$

In particular, if $k$ satisfies some sort of slow oscillation condition, such as

$$\lim_{x\to\infty}\int_x^{x+1}|k(y)-k(x)|\,dy \;=\; 0, \tag{21.4.13}$$

then $\lim_{x\to\infty} k(x) = A$; see Exercise 10.

Wiener showed that Theorem 21.4.6 could be used to derive Littlewood's theorem, Theorem 21.2.2, and Ikehara's theorem, Theorem 21.3.2. We note that one of the transformations that Wiener used was (21.3.12). The specific function $f$ in (21.3.12) has a nowhere vanishing Fourier transform; see Exercise 8 (d).

## 21.5 A theorem of Malliavin and applications

The tauberian theorems considered above give asymptotic results, but with no estimate of the rate of convergence. An example of a result with an error estimate, obtained by taking advantage of information in the complex plane, is a theorem of Malliavin [95]. The setting of the theorem is a transform known (up to a possible sign change) as the *Stieltjes transform*

$$\alpha(s) \;\to\; \int_0^\infty \frac{d\alpha(\lambda)}{\lambda - z} \;=\; \int_0^\infty f_y(x-\lambda)\,d\alpha(\lambda),$$

where $z = x+iy$ and $f_y(x) = -1/(x+iy)$. If $\alpha'(t) = k(t)$, this has the form (21.4.10), but with an extra parameter that allows for more information.

Malliavin's result was generalized by Pleijel [116]. The (relatively) simple proof here is due to Pleijel.

**Theorem 21.5.1.** (Malliavin) *Suppose that*

$$f(z) \;=\; \int_0^\infty \frac{d\sigma(\lambda)}{\lambda - z} \tag{21.5.1}$$

*satisfies the estimate*

$$f(z) \;=\; a(-z)^\alpha + O(z^\beta), \quad -1<\alpha<0,\; \beta<\alpha, \tag{21.5.2}$$

*as $z\to\infty$ along the curve $L \;=\; \{t\pm it^\gamma,\; t\geq 0\}$, where $0\leq\gamma<1$. Then as $X\to\infty$,*

$$\sigma(X)-\sigma(0) \;=\; aX^{\alpha+1}\frac{\sin\pi(\alpha+1)}{\pi(\alpha+1)} + O(X^{\alpha+\gamma}) + O(X^{\beta+1}) + A, \tag{21.5.3}$$

*where $O(X^{\beta+1})$ should be replaced by $O(\log X)$ if $\beta=-1$.*

*Proof:* We may modify the curve $L$ near the origin so that $L$ avoids the line of integration $\{\lambda : \lambda\geq 0\}$. For large $X$, let $L(Z)$ denote the portion of the curve starting at $\bar{Z} = X - iy$ and ending at $Z = X+iy$, $Y = X^\gamma$. Note that the assumptions imply that

$$Y\,f(Z) \;=\; O(X^{\alpha+\gamma}). \tag{21.5.4}$$

Now

$$
\begin{aligned}
I(Z) &\equiv \frac{1}{2\pi i}\int_{L(Z)} f(z)\,dz \qquad (21.5.5)\\
&= \frac{1}{2\pi i}\int_{L(Z)} [f(z)-a(-z)^{\alpha}]\,dz + \frac{a}{2\pi i}\int_{L(Z)} (-z)^{\alpha}\,dz\\
&= I_1(Z)+I_2(Z).
\end{aligned}
$$

As $Z\to\infty$,

$$
I_1(Z) \;=\; \begin{cases} O(X^{\beta+1}) & \text{if } -1<\beta<0,\\ O(\log X) & \text{if } \beta=-1,\\ A+O(X^{\beta+1}) & \text{if } \beta<-1. \end{cases} \qquad (21.5.6)
$$

The second integral $I_2(Z)$ is

$$
\begin{aligned}
\frac{a}{2\pi i(\alpha+1)}(-1)^{\alpha}[Z^{\alpha+1}-\bar{Z}^{\alpha+1}] &= \frac{a}{\pi(\alpha+1)}|Z|^{\alpha+1}\sin[(\alpha+1)(\pi-\arg Z)].\\
&= aX^{\alpha+1}\frac{\sin\pi(\alpha+1)}{\pi(\alpha+1)}+O(X^{\alpha+\gamma}). \qquad (21.5.7)
\end{aligned}
$$

To complete the proof we need to show that

$$
I(Z)-\sigma(X)+\sigma(0) \;=\; O(X^{\alpha+\gamma}). \qquad (21.5.8)
$$

If we insert into (21.5.5) the definition of $f$ and change the order of integration, we find

$$
I(Z) \;=\; \frac{1}{\pi}\int_0^{\infty} \nu(Z,\lambda)\,d\sigma(\lambda), \qquad (21.5.9)
$$

where $\nu(Z,\lambda)$ is the angle between the negative real direction and the direction from $\lambda$ to $Z$; see Figure 21.2. This means that

$$
\lambda - Z \;=\; |\lambda - Z|\,e^{-i\nu}
$$

so

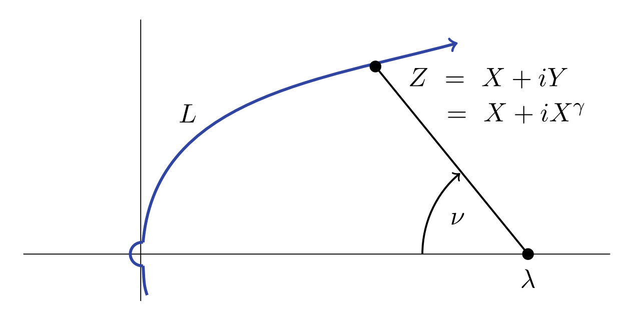



**Fig. 21.2** The angle $\nu$

$$\frac{Y}{\lambda - Z} = \frac{Y}{|\lambda - Z|}e^{i\nu} = \sin\nu(\cos\nu + i\sin\nu) = \tfrac{1}{2}\sin(2\nu) + i\sin^2\nu.$$

It follows that

$$Yf(Z) = \frac{1}{2}\int_0^\infty \sin 2\nu\, d\sigma(\lambda) + i\int_0^\infty \sin^2\nu\, d\sigma(\lambda). \tag{21.5.10}$$

Note also that $\lambda < X$ if and only if $\nu > \pi/2$, so for some $c > 0$,

$$\begin{aligned} \lambda < X &\quad\Rightarrow\quad |\nu - \pi - \tfrac{1}{2}\sin 2\nu| \le c\sin^2\nu; \\ X < \lambda &\quad\Rightarrow\quad |\nu - \frac{1}{2}\sin 2\nu| \le c\sin^2\nu. \end{aligned}$$

Therefore

$$\begin{aligned} I(Z) - \frac{1}{\pi}Y\,\mathrm{Re}\,f(Z) &= \frac{1}{\pi}\int_0^\infty (\nu - \tfrac{1}{2}\sin 2\nu)\, d\sigma(\lambda) \\ &= \int_0^X d\sigma(\lambda) + \frac{1}{\pi}\int_0^X (\nu - \pi - \tfrac{1}{2}\sin 2\nu)\, d\sigma(\lambda) + \frac{1}{\pi}\int_X^\infty (\nu - \tfrac{1}{2}\sin 2\nu)\, d\sigma(\lambda) \\ &= \sigma(X) - \sigma(0) + \frac{1}{\pi}\int_0^X (\nu - \pi - \tfrac{1}{2}\sin 2\nu)\, d\sigma(\lambda) + \frac{1}{\pi}\int_X^\infty (\nu - \tfrac{1}{2}\sin 2\nu)\, d\sigma(\lambda). \end{aligned}$$

Combining this with (21.5.10) and (21.5.11),

$$|I(Z) - \sigma(X) + \sigma(0)| \le \frac{1}{\pi}|Y\,\mathrm{Re}\,f(Z)| + \frac{c}{\pi}|Y\,\mathrm{Im}\,f(Z)|.$$

Together with (21.5.4), this inequality implies (21.5.8). □

One might well ask: under what circumstances would one have the kind of information (21.5.2) along a curve in $\mathbb{C}$? And what use would one make of it? Here is a brief description. Suppose that $M$ is a compact manifold with a Riemannian metric and $A$ is the corresponding Laplace–Beltrami operator. The easiest example is the $n$-torus

$$T^n = \{(\omega_1, \omega_1, \ldots, \omega_n)\}, \quad \omega_j \in \mathbb{C},\ |\omega_j| = 1.$$

This can be coordinatized with $\omega_j = e^{ix_j}$, and the Laplace–Beltrami operator for the flat metric is

$$A = -\left(\frac{\partial^2}{\partial {x_1}^2} + \frac{\partial^2}{\partial {x_2}^2} + \cdots + \frac{\partial^2}{\partial {x_n}^2}\right).$$

For this example it is easy to find the eigenvalues, i.e. the constants $\lambda_j$ such that $Au = \lambda u$ has a non-zero solution $u(x_1, \ldots, x_n)$. In the general case one can show that the eigenvalues are positive and that the associated $L^2$ space has an orthonormal basis consisting of eigenvectors:

$$A\varphi_n = \lambda_n\varphi_n, \quad 0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n \ldots .$$

For $z \in \mathbb{C}\setminus\mathbb{R}$, the solution of $(A - z)u = f$ is given by a resolvent kernel $R_z$:

$$u(x) \;=\; \int_M R_z(x,y)\, f(y)\, dm(y).$$

Expanding in normalized eigenfunctions, one finds that

$$R_z(x,y) \;=\; \sum(\lambda_n - z)^{-1} \varphi_n(x)\varphi_n(y).$$

In particular

$$\int_M R_z(x,x)\, dm(x) \;=\; (\lambda - z)^{-1} \sum \lambda_n.$$

Setting $n(\lambda)$ equal to the number of eigenvalues (counting multiplicity) $\leq \lambda$, we may write the last sum as a Stieltjes integral:

$$\int_0^\infty \frac{1}{\lambda - z}\, dn(\lambda).$$

It is possible to get fairly detailed asymptotic information about the resolvent kernel $R_z$, and therefore to deduce information about the asymptotic distribution of the eigenvalues $\{\lambda_n\}$, of the form

$$n(\lambda) \;\sim\; a\lambda^{n/2} + O(\lambda^{(n-1)/2}).$$

For this and other references, see Agmon [3]

## Exercises

1. Show that (21.0.3) implies (21.0.4).
2. Suppose that the sequence $\{s_n\}$ is non-decreasing. Show that (21.0.4) implies (21.0.3).
3. (a) Consider the example (21.1.7). Show that $f = g'$ satisfies $f' = O(x^{\varepsilon - 1})$ and that (21.1.1) is satisfied with $A = 0$, but (21.1.3) is not satisfied.

   (b) Show that for *any* decay condition weaker than (21.1.2), i.e. a condition $f'(x) = O[\eta(x)]$, where $\eta(x)/x \to \infty$ as $x \to \infty$, there is an example for which (21.1.1) holds but (21.1.3) does not.
4. Prove (21.3.3).
5. Show that Karamata's theorem (say with $a = 1$) can fail if $\alpha$ is not monotone.
6. (a) Translate Ikehara's theorem to a particular case of Karamata's theorem with $a = 1$.

   (b) Show that the conclusion from Ikehara's theorem implies the conclusion of the corresponding case of Karamata's theorem.
7. Verify that $U(\xi/2\varepsilon)$ is the Fourier transform of the function (21.4.7).
8. Show that each of the following functions has a nowhere vanishing Fourier transform.

   (a) $f(x) \;=\; e^{-|x|}$.

(b) $f(x) = 0, \ x<0, f(x) = e^{-x}, \ x \geq 0.$

(c) $f(x) = e^{-x^2/2}.$

(d) $f(x) = e^x \exp(-e^x).$

9. (a) Suppose that $g$ belongs to $L^1$, $k$ is bounded, and $\lim_{x\to+\infty} k(x) = A$. Prove that

$$\lim_{x\to+\infty} \int_{-\infty}^{\infty} g(x-y)\,k(y)\,dy = A\int_{-\infty}^{\infty} g(y)\,dy.$$

(b) Deduce the non-oscillatory version of Theorem 21.4.6: if $k$ is bounded and non-decreasing, then (21.4.10) implies (21.4.11).

10. Suppose that $f$ and $k$ satisfy the hypotheses of Theorem 21.4.6.
Show that if $k$ satisfies (21.4.13), then $k$ has limit $A$ as $x \to \infty$.

11. The next exercises present generalizations of the discrete and continuous versions of Hardy's theorem in Section 21.1.
The *fractional integral* of order $\alpha > 0$ of a continuous function $f : [0,\infty) \to \mathbb{C}$ is defined for $x \geq 0$ by

$$I_\alpha f(x) = \frac{1}{\Gamma(\alpha)} \int_0^x (x-y)^{\alpha-1} f(y)\,dy$$

A related averaging method is

$$J_a f(x) = \frac{\Gamma(a+1)}{x^\alpha} I_\alpha f(x) = \frac{\alpha}{x^\alpha} \int_0^x (x-y)^{\alpha-1} f(y)\,dy.$$

Prove that if $\lim_{x\to\infty} f(x) = A$, then $\lim_{x\to\infty} J_\alpha f(x) = A$.

12. Conversely, let $g(x) = x^\alpha J_\alpha f(x)$ and suppose that $\lim_{x\to\infty} J_\alpha f(x) = A$. Suppose also that $f$ satisfies

$$|f(x+h) - f(x)| \leq K\frac{h^\alpha}{x^\alpha}, \quad h \geq h_0. \tag{21.5.11}$$

Let $g = x^\alpha J_\alpha f$ and use the assumption on the limit of $J_\alpha f$ to estimate

$$\frac{g(x+h)-g(x)}{h^\alpha} - \frac{A}{\Gamma(x+1)}.$$

Use (21.5.11) to estimate

$$\frac{g(x+h)-g(x)}{h^\alpha} - \frac{f(x)}{\Gamma(\alpha+1)}.$$

13. Use the previous exercise to show that the assumption $\lim J_a f(x) = A$ and the tauberian condition (21.5.11) imply that $\lim_{x\to\infty} f(x) = A$. Note that the case $\alpha = 1$ is Hardy's theorem as in Section 21.1.

14. Given a real or complex sequence $\mathbf{a} = \{a_n\}_{n=1}^{\infty}$ and an index $\alpha > 0$, define the average sequence $\mathbf{a}^{(\alpha)}$ to be the sequence with terms

$$a_n^{(\alpha)} = \frac{\alpha}{n} \sum_{k=1}^{n} \left(1 - \frac{k}{n}\right)^{\alpha-1} a_k.$$

(These are closely related to Cesàro $(C, \alpha)$ means.) Use Exercise 13 or an analogous argument to show that if **a** has limit $A$, then so does $\mathbf{a}^{(\alpha)}$.

15. Use Exercise 13 or an analogous argument to show that if $\mathbf{a}^{(\alpha)}$ has limit $A$ and

$$|a_{n+k} - a_n| \le K \frac{k^\alpha}{n^\alpha}, \quad k > 0,$$

then **a** has limit $A$. With $\alpha = 1$, this is Hardy's tauberian theorem for series.

## Remarks and further reading

For more on the history of the subject and further developments, see Wiener [143], Korevaar [77], and Choimet and Queféllec [33].

# Chapter 22
# Asymptotics and the method of steepest descent



In both pure and applied mathematics, a number of questions lead to estimating integrals of the type

$$\int_C g(z)\, e^{\lambda f(z)}\, dz \tag{22.0.1}$$

as the parameter $\lambda$ becomes large. Here we assume that $g$ and $f$ are holomorphic in some domain $\Omega$, and $C$ is a curve subject to some constraints. Typically the endpoints of $C$ are fixed (the positions possibly depending on $\lambda$). Moreover, the domain $\Omega$ is usually simply connected, so, given the endpoints, the value of the integral is independent of the choice of the curve $C$. The idea is to choose a curve that makes the estimation as easy as possible.

In this chapter we describe the general strategy known as the "method of steepest descent" and then apply this method to two examples: the asymptotics of the Airy integral, and the Hardy–Ramanujan asymptotics for the number-theoretic partition function.

## 22.1 The method of steepest descent

Consider the integral (22.0.1) with fixed endpoints (possibly at infinity). Assume for now that the parameter $\lambda$ is real and positive. Let $f(x+iy) = u(x,y) + iv(x,y)$, where $u$ and $v$ are real, and assume that $f$ is not constant. (We shall also abuse notation and write $u(z)$, $v(z)$ as convenient.) As the parameter $\lambda$ in (22.0.1) becomes large, it is clear that the main contribution to the integral will come from the part of the curve where $u$ is largest.

If there is a curve in the admissible domain $\Omega$ such that the maximum value of $u$ occurs at an endpoint, then only behavior at or near that endpoint is relevant to the asymptotics. Assume, instead, that on each admissible curve $u$ reaches a maximum value greater than the value at the endpoints. Suppose that $C$ is such a curve, and that the maximum value is attained at a point $z_0$. If the gradient of $u$ is not zero at



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_22

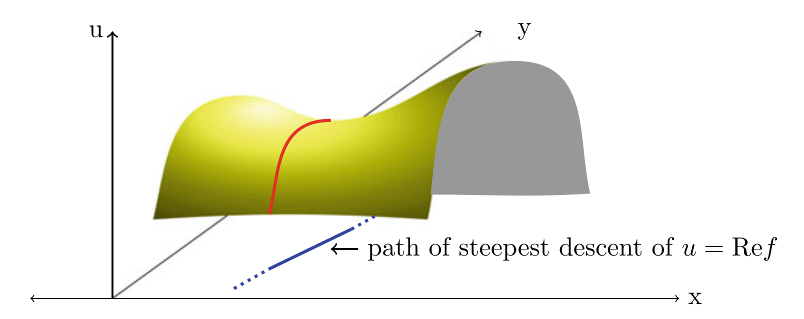



**Fig. 22.1** Steepest descent path

$z_0$, then by displacing the curve slightly to the left or right, we can reduce the value of $u$ on a stretch of the curve. Thus what we need to examine is the case that the maximum of $u$ on the curve occurs at a critical point $z_0$: $u_x(z_0) = u_y(z_0) = 0$. The Cauchy–Riemann equations imply that also $v_y(z_0) = v_x(z_0) = 0$, so $z_0$ is a zero of $f'$. Note that $u$ cannot have a local maximum at $z_0$, since then $|\exp(f)|$ would have a maximum, contradicting the strong maximum principle (Corollary 4.2.2). Therefore $z_0$ is a saddle point of some sort for the surface $\{z, u(z))\} \in \mathbb{C} \times \mathbb{R} \equiv \mathbb{R}^3$.

Let us assume that at each critical point $z_0$ that needs to be considered, $f''(z_0) \neq 0$. Then, up to a translation and rotation, the surface looks, locally, like the one pictured in Figure 22.1. There are two paths on the surface that follow the gradient of $u$ and pass through $u(z_0)$. On one path, $u$ is increasing most rapidly, and on the other path $u$ is decreasing most rapidly, as $z$ moves away from $z_0$. The Cauchy–Riemann equations imply that the gradient of $u$ is orthogonal to the gradient of $v$, so $v$ is constant along these paths. In particular, near a critical point on a curve $C$, the curve can be adjusted to follow (at least for a while) the path of steepest descent, and the principal contribution to the integral (22.0.1) along the curve will come from such portions of the curve.

In view of this discussion, the problem comes down to examining an integral (22.0.1), where $C$ now denotes a portion of the path of steepest descent near a critical point $u(z_0)$. Near $z_0$,

$$f(z) \;=\; f(z_0) - (z - z_0)^2 f_1(z) \;=\; f(z_0) - (z - z_0)^2 a\,[1 + O(z - z_0)]$$

and, by assumption, $a = f_1(z_0) = -f''(z_0)/2 \neq 0$. The imaginary part of $f$ is constant along $C$ and the real part is less than $\text{Re}\, f(z_0)$ except at $z = z_0$. Therefore $(z - z_0)^2 f_1(z)$ is positive except at $z_0$. We make a determination of $a^{1/2}$, and use the principal square root of the term in braces to write

$$f(z) \;=\; f(z_0) - t(z)^2, \qquad t(z) \;=\; (z - z_0)\, a^{1/2}[1 + O(z - z_0)]^{1/2},$$

with $t^2 > 0$ on $C$ except at $z = z_0$. By the inverse function theorem, Theorem 1.3.8, $z$ can be obtained as a holomorphic function of $t$ near $z = z_0$:

$$z = w(t).$$

Then the integral we are considering can be written

$$I(\lambda) = e^{\lambda f(z_0)} \int_{-\delta}^{\delta} e^{-\lambda t^2} g(w(t)) w'(t)\, dt, \qquad (22.1.1)$$

for some $\delta > 0$.

We need two more observations. First, $(g \circ w)w'$ can be expanded in a power series, and terms that vanish more rapidly at $z = z_0$ will contribute less to the integral. Therefore we may obtain an asymptotic expansion by integrating term-by-term. Second, each such term of the integral in (22.1.1) has the form

$$\int_{-\delta}^{\delta} e^{-\lambda t^2} c_n t^n\, dt.$$

It is easily seen that, modulo an error term that is $O(e^{-\lambda \delta^2})$, this integral can be replaced by

$$\int_{-\infty}^{\infty} e^{-\lambda t^2} c_n t^n\, dt.$$

The integrand is odd, so the result is 0 if $n$ is odd. Otherwise we obtain

$$\begin{aligned} 2c_n \int_0^{\infty} e^{-\lambda t^2} t^n\, dt &= c_n \lambda^{-(n+1)/2} \int_0^{\infty} e^{-s} s^{(n-1)/2}\, ds \\ &= c_n \lambda^{-(n+1)/2} \Gamma\left(\frac{n+1}{2}\right), \quad \text{if } n \text{ is even.} \qquad (22.1.2) \end{aligned}$$

(See (10.2.2) for the evaluation of the integral.) If $g(z_0) \neq 0$, then the leading coefficient is $g(z_0)/[-f''(0)/2]^{1/2}$. Since $\Gamma(\frac{1}{2}) = \pi^{1/2}$, the leading term of the expansion in this case is

$$I(\lambda) \sim g(z_0) \left[ \frac{2\pi}{-\lambda f''(z_0)} \right]^{1/2} e^{\lambda f(z_0)}. \qquad (22.1.3)$$

## 22.2 The Airy integral

The Airy integral is

$$\mathrm{Ai}\,(z) = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \exp\left( \frac{w^3}{3} - zw \right) dw.$$

The path of integration can be modified so that, for large $w$, it lies in the part of the domain $\{\mathrm{Re}\, w^3 < 0\}$ that adjoins the imaginary axis:

$$\begin{cases} -\frac{\pi}{2} < \arg w < -\frac{\pi}{2} & \text{if } \operatorname{Im} w \ll 0; \\ \frac{\pi}{6} < \arg w < \frac{\pi}{2} & \text{if } \operatorname{Im} w \gg 0. \end{cases} \tag{22.2.1}$$

(See Exercise 1.) Thus we consider

$$\operatorname{Ai}(z) = \frac{1}{2\pi i}\int_C \exp\left(\frac{w^3}{3} - zw\right) dw, \tag{22.2.2}$$

where the curve $C$ lies in the domain (22.2.1). The function so defined is entire in $z$. At the start, we assume $z > 0$ and consider the asymptotics of $\operatorname{Ai}(z)$ as $z \to +\infty$. The integral can be put into the form (22.0.1) by a change of scale: let $\omega = z^{-1/2}w$. Since we are assuming here that $z > 0$, the curve followed by $\omega$ will lie in the same domain (22.2.1). The integral now has the form

$$\begin{aligned} \operatorname{Ai}(z) &= \frac{z^{1/2}}{2\pi i}\int_C \exp\left\{z^{3/2}\left(\frac{\omega^3}{3} - \omega\right)\right\} d\omega \\ &= \frac{z^{1/2}}{2\pi i}\int_C e^{\lambda f(\omega)}\, d\omega, \end{aligned} \tag{22.2.3}$$

with

$$\lambda = z^{3/2}, \qquad f(\omega) = \frac{\omega^3}{3} - \omega.$$

The critical points of the integrand occur where $0 = f'(\omega) = \omega^2 - 1$, i.e. at $\omega = \pm 1$. Now

$$f(\sigma + i\tau) = \left(\frac{\sigma^3}{3} - \sigma\tau^2 - \sigma\right) + i\left(\sigma^2\tau - \tau - \frac{\tau^3}{3}\right).$$

Then $\operatorname{Im} f(\pm 1) = 0$. The curves with imaginary part zero that pass through $\pm 1$ are determined by $\tau = 0$ or $\frac{1}{3}\tau^2 = \sigma^2 - 1$. Since $u(\sigma, 0) = \frac{1}{3}\sigma^3 - \sigma$ has a local maximum at $\sigma = -1$ and a local minimum at $\sigma = 1$, it follows that the branch of the hyperbola $\tau^2 = 3(\sigma^2 - 1)$ is a path of steepest descent from the critical point at $\omega = 1$. Moreover, this path lies in the domain (22.2.1); see Figure 22.2.

Following the prescription in Section 22.1, we write

$$\begin{aligned} f(\omega) &= f(1) + \frac{f''(1)}{2}(\omega - 1)^2 + \frac{f'''(1)}{6}(\omega - 1)^3 \\ &= -\frac{2}{3} - (\omega - 1)^2(-1)[1 + \tfrac{1}{3}(\omega - 1)] \\ &= -\frac{2}{3} - t^2(\omega), \qquad t^{\pm}(\omega) = \mp i(\omega - 1)\left[1 + \frac{1}{6}(\omega - 1) + \dots\right]. \end{aligned}$$

In this case, $t$ may be defined globally along the curve, since $\operatorname{Re}(\omega - 1)$ stays positive. Note that for $\omega$ in the upper half of the curve $C$, we have $d\omega/dt = 1/(dt^+/d\omega) = i$, and $d\bar{\omega}/dt = 1/(dt^-/d\omega) = -i$ at $\omega = 1$ (i.e. at $t = 0$). Thus our integral is

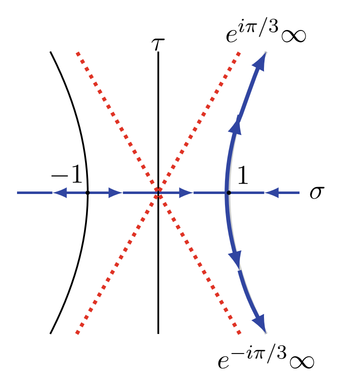



**Fig. 22.2** Steepest path for $\mathrm{Ai}(z)$

$$\frac{z^{1/2}}{2\pi i} e^{-\frac{2}{3}\lambda} \int_{-\infty}^{\infty} \left( \frac{d\omega}{dt} - \frac{d\bar{\omega}}{dt} \right) e^{-\lambda t^2} \, dt \;=\; \frac{e^{-\frac{2}{3}z^{3/2}}}{2\pi^{1/2} z^{1/4}} \left[ 1 + O(z^{-3/2}) \right]. \tag{22.2.4}$$

We assumed up to this point that $z$ was positive. The steepest descent curve is asymptotic to the rays that bisect the domain (22.2.1). It follows that $w = \omega z^{1/2}$ belongs to the domain (22.2.1) so long as

$$|\arg z| \;\leq\; \frac{\pi}{3} - \delta, \quad \delta > 0. \tag{22.2.5}$$

Therefore the estimate (22.2.4) is valid uniformly for $z$ in the sector (22.2.5).

## 22.3 The partition function and the Hardy–Ramanujan formula

If $n$ is a positive integer, the *partition function* $p(n)$ is the number of ways of writing $n$ as a sum of positive integers in non-increasing order. By convention, $p(0) = 1$ (see below for an explanation). Clearly $p(1) = 1$, while

$$2 = 2 = 1+1, \qquad 3 = 3 = 2+1 = 1+1+1,$$
$$4 = 4 = 3+1 = 2+2 = 2+1+1 = 1+1+1+1,$$

so $p(2) = 2$, $p(3) = 3$, $p(4) = 5$. It can easily be checked that $p(5) = 7$, $p(6) = 11$, $p(7) = 15$, and $p(8) = 22$. This gives some hint of how rapidly $p(n)$ grows. Our goal is to derive a result of Hardy and Ramanujan:

$$p(n) \;\sim\; \frac{1}{4n\sqrt{3}} \exp\left( \sqrt{\frac{2}{3}}\, \pi n^{1/2} \right) \tag{22.3.1}$$

as $n \to \infty$.

Euler noted that for $|w| < 1$,

$$\prod_{m=1}^{\infty} \frac{1}{1-w^m} = \sum_{n=0}^{\infty} p(n) w^n; \tag{22.3.2}$$

see Exercise 3. (This explains the convention that $p(0) = 1$.) We can pick out the coefficient $p(n)$ by using the fact that

$$\frac{1}{2\pi i} \int_{|w|=r} w^{n-m} \frac{dw}{w} = \begin{cases} 1 & \text{if } n = m, \\ 0 & \text{if } n \neq m. \end{cases} \tag{22.3.3}$$

It will be convenient to change variables here. Letting $w = e^{2\pi i z}$, $\operatorname{Im} z > 0$, we have

$$f(z) \equiv \prod_{m=1}^{\infty} \frac{1}{1-e^{2\pi i m z}} = \sum_{n=0}^{\infty} p(n) e^{2n\pi i z}. \tag{22.3.4}$$

Let $L$ denote the line segment

$$L = \{z : z = x + i\delta, \ -\tfrac{1}{2} \le x \le \tfrac{1}{2}\},$$

where $0 < \delta < 1$. As $z$ runs along $L$, $w = e^{2\pi i z}$ runs around the circle $|w| = e^{-2\pi\delta}$. Therefore picking out $p(n)$ by using (22.3.3) with (22.3.4) gives

$$p(n) = \int_L f(z) \, e^{-2n\pi i z} \, dz. \tag{22.3.5}$$

A key role in the analysis of $f$ is played by the functional equation for $f$. The proof is given in Section 22.4.

**Theorem 22.3.1.** *The function $f$ satisfies the functional equation*

$$f(z) = \sqrt{\frac{z}{i}} \exp\left(\frac{i\pi}{12}\left[z + \frac{1}{z}\right]\right) f\left(-\frac{1}{z}\right). \tag{22.3.6}$$

Taking the logarithm of (22.3.4) and using (1.5.2), we find that

$$f(z) = 1 + O(e^{-2\pi \operatorname{Im} z}) \quad \text{if } \operatorname{Im} z \ge 1. \tag{22.3.7}$$

Applying this to $f(-1/z)$, and taking into account the functional equation (22.3.6), suggests writing

$$p(n) = p_1(n) + E(n) = \int_L \sqrt{\frac{z}{i}} \, e^{i\pi(z+1/z)/12} \, e^{-2n\pi i z} \, dz + E(n). \tag{22.3.8}$$

Let us estimate the remainder term

$$E(n) = \int_L \sqrt{\frac{z}{i}} \, e^{i\pi(z+1/z)/12} \, [f(-1/z) - 1] \, e^{-2n\pi i z} \, dz. \tag{22.3.9}$$

For $z = x + i\delta$ in $L$, $\operatorname{Im} z = \delta$ and $\operatorname{Im}(-1/z) = \delta/(\delta^2 + x^2)$. Thus

$$\left|\sqrt{\frac{z}{i}}\, e^{i\pi(z+1/z)/12} e^{-2n\pi iz}\right| = |z|^{1/2} e^{\pi[-\delta+\delta/(\delta^2+x^2)]/12} e^{2n\pi\delta}$$
$$= O\left(e^{\pi\delta/[12(\delta^2+x^2)]} e^{2n\pi\delta}\right). \tag{22.3.10}$$

For $x$ close to 0, $\delta/(\delta^2 + x^2) \geq 1$. When this is the case, (22.3.7) gives

$$|f(-1/z) - 1| = O(e^{-2\pi\delta/(\delta^2+x^2)}).$$

Combining this with (22.3.9) we see that the part of the integral (22.3.9) where $\operatorname{Im}(-1/z) \geq 1$ is $O(e^{2n\pi\delta})$.

For $z = x + i\delta$ in $L$, with $x$ close to $\pm 1/2$, it follows that $\delta/(\delta^2 + x^2) \leq 1$. To estimate the part of the integral (22.3.9) in this range we note first that with $\operatorname{Im} z > 0$ we have $|1 - e^{2n\pi iz}| \geq |1 - e^{-2n\pi \operatorname{Im} z}|$, so

$$|f(z)| \leq f(i \operatorname{Im} z). \tag{22.3.11}$$

Suppose that $z = x + iy$, $0 < y \leq 1$. Then $f(-1/iy) = f(i/y)$, so (22.3.7) implies that $f(i/y) = O(1)$. Therefore (22.3.6) implies

$$|f(z)| \leq f(iy) = O(e^{\pi/12y}), \qquad y = \operatorname{Im} z \leq 1.$$

Applying this to $f(-1/z)$ for $z \in L$, we obtain

$$|f(-1/z)| = O\left(e^{\pi(\delta^2+x^2)/12\delta}\right) = O\left(e^{\pi/48\delta}\right), \tag{22.3.12}$$

if $\delta/(\delta^2 + x^2) \leq 1$. The same estimate applies, in this case, to $|f(-1/z) - 1|$. Taking into account (22.3.10), we find that the part of the integral (22.3.9) where $\operatorname{Im}(-1/z) \leq 1$ is $O(e^{\pi/48\delta} e^{2n\pi\delta})$. Altogether, then,

$$|E(n)| = O(e^{\pi/48\delta} e^{2n\pi\delta}).$$

This estimate is optimal when $2\pi n\delta + \pi/(48\delta)$ is minimal, which occurs for $\delta = \delta_n = 1/4\sqrt{6n}$ and gives the estimate

$$|E(n)| = O(e^{Kn^{1/2}/2}), \qquad K = \pi\sqrt{\frac{2}{3}}. \tag{22.3.13}$$

It remains to estimate the integral defining $p_1(n)$ in (22.3.8). The proof presented here is based on the presentation in Stein and Shakarchi [132].

**Theorem 22.3.2.** (Hardy–Ramanujan) *If $p(n)$ denotes the partition function, then*

*(i)* $p(n) \sim \dfrac{1}{4n\sqrt{3}}\, e^{Kn^{1/2}}$ *as $n \to \infty$, where* $K = \pi\sqrt{\frac{2}{3}}$.

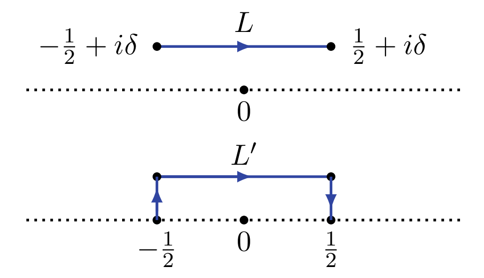



**Fig. 22.3** Contours $L$ and $L'$

*(ii) More precisely*

$$p(n) = \frac{1}{2\pi\sqrt{2}}\frac{d}{dn}\left\{\frac{e^{K(n-\frac{1}{24})^{1/2}}}{(n-\frac{1}{24})^{1/2}}\right\} + O(e^{\frac{K}{2}n^{1/2}}).$$

*Proof:* First we modify the contour $L$ in Figure 22.3 by adding to it two small vertical line segments that join $-\frac{1}{2}$ to $-\frac{1}{2}+i\delta$ and $\frac{1}{2}+i\delta$ to $\frac{1}{2}$. We label the new contour $L'$; see Figure 22.3.

Note that $\sqrt{z/i}\,e^{i\pi/12z}$ is $O(1)$ on the two added segments. Therefore for the integral $p_1(n)$ in (22.3.8), the addition contributes

$$O(e^{2\pi n\delta}) = O(e^{2\pi n^{1/2}/4\sqrt{6}}) = O(e^{Kn^{1/2}/4}),$$

which is smaller than the allowed error. Therefore, we may incorporate this contribution into the error term $E(n)$. Without introducing further notation we will rewrite $p_1(n)$ with $L$ replaced by $L'$ in the integral defining $p_1(n)$, namely,

$$p_1(n) = \int_{L'}\sqrt{\frac{z}{i}}\,e^{i\pi(z+1/z)/12}\,e^{-2\pi inz}dz. \tag{22.3.14}$$

Next we make a change of variables $z=\mu z'$ so that the exponential functions are combined into one of the form

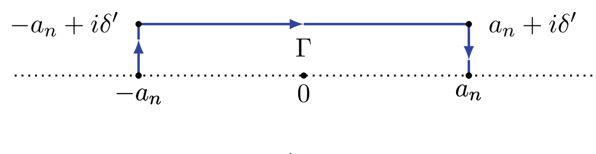
$$\exp\left\{iA\left(\frac{1}{z}-z\right)\right\}.$$

This can be achieved by choosing

$$A = \frac{\pi}{\sqrt{6}}\left(n-\frac{1}{24}\right)^{1/2} \qquad \text{and} \qquad \mu = \frac{1}{2\sqrt{6}}\left(n-\frac{1}{24}\right)^{-1/2}. \tag{22.3.15}$$

The result is

$$p_1(n) = \mu^{3/2}\int_{\Gamma} e^{-\lambda F(z')}\sqrt{\frac{z'}{i}}\,dz', \tag{22.3.16}$$



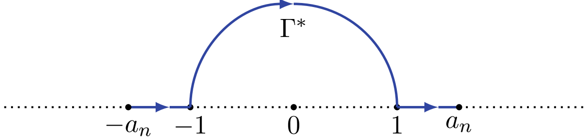




**Fig. 22.4** Contours $\Gamma$ and $\Gamma^*$

where

$$F(z') = i(z' - \frac{1}{z'}), \qquad \lambda = \frac{\pi}{\sqrt{6}}\left(n - \frac{1}{24}\right)^{1/2}, \tag{22.3.17}$$

and the contour $\Gamma = \mu^{-1}L'$ is now the union of three segments $[-a_n, -a_n + i\delta']$, $[-a_n + i\delta', a_n + i\delta']$ and $[a_n + i\delta', a_n]$. Here $a_n = \frac{1}{2}\mu^{-1} = \sqrt{6}\left(n - \frac{1}{24}\right)^{1/2} \approx \sqrt{6n}$, while $\delta' = \delta\mu^{-1} = 2\sqrt{6}\ \left(n - \frac{1}{24}\right)^{1/2}/4\sqrt{6n} \sim 1/2$, as $n \to \infty$.

We apply the method of steepest descent to the integral (22.3.16). Note that the *phase function* $F(z) = i\left(z - \frac{1}{z}\right)$ has only one critical point $z = i$ in the upper half plane. Furthermore, the two curves passing through $i$ on which $F$ is real are the imaginary axis and the unit circle. On the imaginary axis, $F$ has a maximum at $z = i$. On the unit circle, $F$ has a minimum at $z = i$. Thus we deform the contour $\Gamma$ into the contour $\Gamma^*$, which consists of the segments $[-a_n, -1]$ and $[1, a_n]$, together with the upper semicircle joining $-1$ and 1; see Figure 22.4.

Then

$$p_1(n) = \mu^{3/2} \int_{\Gamma^*} e^{-sF(z)} \sqrt{\frac{z}{i}}\, dz.$$

The contributions from the integral on the segments $[-a_n, -1]$ and $[1, a_n]$ are small, because on the real axis the exponential has absolute value 1 and the integrand is bounded by $\sup_{|z| \le a_n} |z|^{1/2}$. This yields an error of the order $O(a_n^{3/2}\mu^{3/2}) = O(1)$.

Finally, we come to the principal part, which is the integral on the semicircle, oriented in the negative direction. Let us write $z = e^{i\theta}$. Then, $i(z - \frac{1}{z}) = -2\sin\theta$ and $dz = ie^{\theta}d\theta$. This gives the contribution

$$-\mu^{3/2} \int_0^{\pi} e^{2s\sin\theta} e^{i3\theta/2} \sqrt{i}\, d\theta$$
$$= \mu^{3/2} \int_{-\pi/2}^{\pi/2} e^{2s\cos\theta} \left(\cos\frac{3\theta}{2} + i\sin\frac{3\theta}{2}\right) d\theta.$$

To apply formula (22.1.3), we take $\lambda = 2s, f(\theta) = \cos\theta$ and $g(\theta) = \cos\frac{3\theta}{2} + i\sin\frac{\theta}{2}$. Then $f(0) = g(0) = 1$ and $f''(0) = -1$. Thus, the above quantity is equal to

$$\mu^{3/2} e^{2s} \frac{\sqrt{2\pi}}{(2s)^{1/2}} \left[1 + O(s^{-1/2})\right].$$

Since $s = \frac{\pi}{\sqrt{6}}(n - \frac{1}{24})^{1/2}$, $\frac{2\pi}{\sqrt{6}} = \pi\sqrt{\frac{2}{3}} = K$ and $\mu = \frac{\sqrt{6}}{12}(n - \frac{1}{24})^{-1/2}$, we obtain

$$p(n) \;=\; \frac{1}{4n\sqrt{3}} e^{Kn^{1/2}} [1 + O(n^{-1/4})]$$

which is the first conclusion of the theorem.

To establish the asymptotic formula in (ii), we retrace some of our earlier steps. With $p_1(n)$ defined by (22.3.14), we write

$$p_1(n) \;=\; \frac{d}{dn} q(n) + e(n), \tag{22.3.18}$$

where

$$q(n) \;=\; \frac{1}{2\pi} \int_{L'} \sqrt{\frac{i}{z}} e^{i\pi(z+1/z)/12} e^{-2\pi inz} dz, \tag{22.3.19}$$

$L'$ is shown in Figure 22.3, and $e(n)$ is the term due to the variation of the contour $L' = L'_n$, when forming the derivative in $n$. (Recall that $\delta$ in the contour $L'$ depends on $n$; see (22.3.13).) Using Leibnitz's formula and estimating the resulting expression as before, it is easily seen that $e(n) = O(e^{2\pi n\delta})$, which we have already seen to be $O(e^{\frac{K}{4} n^{1/2}})$; see the first part of the paragraph containing equation (22.3.14). This error can therefore be subsumed in the error term $E(n)$. To analyze $q(n)$, we again make the change of variable $z \to \mu z$, and then replace the resulting contour $\Gamma$ by $\Gamma^*$. As a result, we obtain

$$q(n) \;=\; \frac{\mu^{1/2}}{2\pi} \int_{\Gamma^*} e^{-\lambda F(z)} (z/i)^{-1/2} dz, \tag{22.3.20}$$

where $F(z) = i(z - \frac{1}{z}), \lambda = \frac{\pi}{\sqrt{6}}(n - \frac{1}{24})^{1/2}$ and $\mu = \frac{1}{2\sqrt{6}}(n - \frac{1}{24})^{-1/2}$; see (22.3.15) and (22.3.17).

Since $F$ is purely imaginary on the real axis, the two segment $[-a_n, -1]$ and $[1, a_n]$ of the contour $\Gamma^*$ contribute only terms of the order $O(a_n^{1/2}\mu^{1/2}) = O(1)$.

As in case (i), the main contribution of (22.3.20) comes from the integral on the semicircle. Setting $z = e^{i\theta}, dz = ie^{i\theta} d\theta$, and $i(z - \frac{1}{z}) = -2\sin\theta$ shows that the integral is equal to

$$\begin{aligned} -\frac{\mu^{1/2}}{2\pi} \int_0^{\pi} e^{2s\sin\theta} e^{i\theta/2} i^{3/2} d\theta \;&=\; \frac{\mu^{1/2}}{2\pi} \int_{-\pi/2}^{\pi/2} e^{2s\cos\theta} \left(\cos\frac{\theta}{2} + i\sin\frac{\theta}{2}\right) d\theta \\ &=\; \frac{\mu^{1/2}}{2\pi} \int_{-\pi/2}^{\pi/2} e^{2s\cos\theta} \cos\frac{\theta}{2} d\theta. \end{aligned}$$

Since $\cos\theta = 1-2(\sin\frac{\theta}{2})^2$, we now set $x=\sin\frac{\theta}{2}$. The above integral becomes

$$\frac{\mu^{1/2}e^{2s}}{\pi}\int_{-\frac{1}{\sqrt{2}}}^{\frac{1}{\sqrt{2}}} e^{-4sx^2}dx.$$

It is easily verified that

$$\begin{aligned}\int_{-\frac{1}{\sqrt{2}}}^{\frac{1}{\sqrt{2}}} e^{-4sx^2}dx &= \int_{-\infty}^{\infty} e^{-4sx^2}dx + O\left(\int_{1/\sqrt{2}}^{\infty} e^{-4sx^2}dx\right)\\ &= \frac{\sqrt{\pi}}{2s^{1/2}} + O(e^{-2s});\end{aligned} \tag{22.3.21}$$

see Exercise 4

Gathering all the error terms together gives

$$p(n) \;=\; \frac{d}{dn}\left\{\mu^{1/2}\frac{e^{2s}}{\pi}\frac{\sqrt{\pi}}{2s^{1/2}}\right\} + O\left(e^{\frac{K}{2}}n^{1/2}\right).$$

Since $s=\frac{\pi}{\sqrt{6}}(n-\frac{1}{24})^{1/2}, \mu=\frac{\sqrt{6}}{12}(n-\frac{1}{24})^{-1/2}$ and $K=\pi\sqrt{\frac{2}{3}}$, the last equation yields

$$p(n) \;=\; \frac{1}{2\pi\sqrt{2}}\frac{d}{dn}\left\{\frac{e^{K(n-\frac{1}{24})^{1/2}}}{(n-\frac{1}{24})^{1/2}}\right\} + O(e^{\frac{K}{2}n^{1/2}}). \qquad \square$$

## 22.4 Proof of the functional equation (22.3.6)

We use an argument due to Siegel [126] in an expanded and more user-friendly form. Recall that the generating function $f$ for the partition function is

$$f(z) \;=\; \prod_{n=1}^{\infty}\frac{1}{1-e^{2\pi inz}}.$$

Note that

$$\frac{q^{2k}}{1-q^{2k}} \;=\; \frac{1}{2}\left[\frac{1+q^{2k}}{1-q^{2k}}-1\right] \;=\; -\frac{1}{2}\,\frac{q^k+q^{-k}}{q^k-q^{-k}}-\frac{1}{2}.$$

If $q \;=\; e^{i\pi z}$, the last expression is

$$\frac{i}{2}\cot k\pi z-\frac{1}{2}.$$

In view of this, consider

$$\begin{aligned}
\log\frac{f(z)}{f(-1/z)} &= \log\prod_{n=1}^{\infty}\frac{1-e^{-2n\pi i/z}}{1-e^{2n\pi iz}}\\
&= \sum_{n=1}^{\infty}[\log(1-e^{-2n\pi i/z})-\log(1-e^{2n\pi iz})]\\
&= \sum_{n=1}^{\infty}\sum_{k=1}^{\infty}\frac{1}{k}\,[e^{2nk\pi iz}-e^{-2nki\pi/z}]\\
&= \sum_{k=1}^{\infty}\frac{1}{k}\left[\frac{e^{2k\pi iz}}{1-e^{2ki\pi z}}-\frac{e^{-2ki\pi/z}}{1-e^{-2ki\pi/z}}\right]\\
&= \frac{i}{2}\sum_{k=1}^{\infty}\frac{1}{k}\,[\cot(k\pi z)+\cot(k\pi/z)]\,.
\end{aligned}$$

For fixed $z\in\mathbb{C}_+$, let

$$g(s) \;=\; \cot s\,\cot(s/z).$$

Given some constant $\nu>0$, the function $h_\nu(s)$ defined by

$$h_\nu(s) \;=\; \frac{g(\nu s)}{s} \;=\; \frac{\cot(\nu s)\,\cot(\nu s/z)}{s}$$

has a triple pole at $s=0$, and two sequences of simple poles. There are poles at $s=\pm k\pi/\nu$ with residues $(1/k\pi)\cot(k\pi/z)$, and simple poles at $s=\pm k\pi z/\nu$ with residues $(1/k\pi)\cot(k\pi z)$, $k=1,2,\dots$. Near $s=0$,

$$\begin{aligned}
g(s) &= \frac{[1-\frac{1}{2}s^2+O(s^4)][1-\frac{1}{2}(s/z)^2+O(s^4)]}{s[1-\frac{1}{6}s^2+O(s^4)](s/z)[1-\frac{1}{6}(s/z)^2+O(s^4)]}\\
&= \frac{z[1-\frac{1}{2}s^2-\frac{1}{2}(s/z)^2+O(s^4)]}{s^2[1-\frac{1}{6}s^2-\frac{1}{6}(s/z)^2+O(s^4)]}\\
&= \frac{z}{s^2}[1-\tfrac{1}{2}s^2-\tfrac{1}{2}(s/z)^2][1+\tfrac{1}{6}s^2+\tfrac{1}{6}(s/z)^2]+O(s^2)\\
&= \frac{z}{s^2}-\frac{z+1/z}{3}+O(s^2).
\end{aligned}$$

It follows that the residue of $h_\nu(s)=g(\nu s)/s$ at the origin is $-(z+z^{-1})/3$.

Let $C$ be the boundary of the parallelogram with vertices $1,z,-1,-z$, oriented with the parallelogram to the left; see Figure 22.5. Let $\nu=(n+\frac{1}{2})\pi$. Then the poles of $h_\nu$ enclosed by $C$ are the origin and the points

$$\pm\frac{k\pi}{n+\frac{1}{2}},\quad \pm\frac{k\pi z}{n+\frac{1}{2}},\quad k=1,2,\dots n,$$

with residues

$$\frac{1}{\pi k}\cot(k\pi z),\qquad \frac{1}{\pi k}\cot(k\pi/z).$$

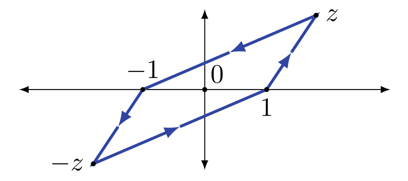



**Fig. 22.5** The contour $C$

Therefore

$$\frac{1}{2\pi i}\int_C h_\nu(s)\,ds = \frac{1}{2\pi i}\int_C \frac{g(\nu s)}{s}\,ds$$
$$= -\tfrac{1}{3}(z+z^{-1}) + \frac{2}{\pi}\sum_{k=1}^{n}\frac{1}{k}[\cot(\pi kz)+\cot(\pi k/z)].$$

As $\nu\to\infty$, $g(\nu s)$ is uniformly bounded on $C$ and converges on the respective sides, minus the corners, to $1,-1,z,-z$; Exercise 5. Therefore

$$\lim_{\nu\to\infty}\int_C \frac{g(\nu s)}{s}\,ds = \int_1^z \frac{ds}{s} - \int_z^{-1}\frac{ds}{s} + \int_{-1}^{-z}\frac{ds}{s} - \int_{-z}^{1}\frac{ds}{s}$$
$$= 4\log z - 2\log(-1) \;=\; 4\log\frac{z}{i}. \tag{22.4.1}$$

Combining these results,

$$\frac{1}{2\pi i}4\cdot\log\frac{z}{i} + \frac{z+z^{-1}}{3} \;=\; \frac{2}{\pi}\sum_{k=1}^{\infty}\frac{1}{k}[\cot(\pi kz)+\cot(\pi k/z)]$$
$$= \frac{4}{i\pi}\log\frac{f(z)}{f(-1/z)},$$

or

$$\frac{1}{2}\log\frac{z}{i} + \frac{i\pi}{12}\left(z+\frac{1}{z}\right) \;=\; \log\frac{f(z)}{f(-1/z)}. \tag{22.4.2}$$

Exponentiating (22.4.2) gives (22.3.6). □

Siegel's argument was designed to prove the functional equation for the *Dedekind eta function*

$$\eta(z) \;=\; e^{i\pi z/12}\prod_{n=1}^{\infty}(1-e^{2n\pi iz}), \quad \operatorname{Im} z>0. \tag{22.4.3}$$

Obviously this is closely related to the function $f$; in fact

$$f(z) \;=\; \frac{e^{i\pi z/12}}{\eta(z)}. \tag{22.4.4}$$

## Exercises

1. Show that the path of the Airy integral can be shifted from the imaginary axis so as to lie in the domain (22.2.1) without changing the value of the integral.
2. Prove that the function $\mathrm{Ai}(z)$ is a solution of the differential equation

$$u''(z) = zu(z).$$

3. Prove the identity (22.3.2)
4. Prove (22.3.21).
5. Prove the limiting relation of $g(\nu s)$ that leads to (22.4.1).
6. The complementary error function $\mathrm{erfc}(z)$ has the contour integral representation

$$\mathrm{erfc}(z) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^{s-2z\sqrt{s}}\,\frac{ds}{s}, \qquad c > 0.$$

   Use the steepest descent method to derive the asymptotic expansion

$$\mathrm{erfc}(z) \sim \frac{e^{-z^2}}{z\sqrt{\pi}}\left(1 - \frac{1}{2z^2} + \frac{3}{4z^4} - \frac{15}{8z^6} + \dots\right),$$

   as $z \to \infty$ in the region $|\arg(z)| \le \frac{3}{4}\pi - \delta$, $\delta > 0$.
7. Hankel's loop integral representation for the reciprocal of the gamma function is

$$\frac{1}{\Gamma(z)} = \frac{1}{2\pi i}\int_{-\infty}^{(0^+)} e^t t^{-z}\,dt,$$

   where the logarithm in $t^{-z} = e^{-t\log z}$ has its principal value. The path of integration starts at $e^{-i\pi}\infty$, goes around the origin once, and ends at $e^{i\pi}\infty$. Use the steepest descent method to show that

$$\frac{1}{\Gamma(z)} \sim e^z z^{-z}\sqrt{\frac{z}{2\pi}}\left(1 - \frac{1}{12z} + \frac{1}{288z^2} - \dots\right)$$

   as $z \to \infty$ in the region $|\arg z| < \frac{1}{2}\pi$.
8. Use the method of steepest descent to derive the asymptotic expansion of the Bessel function

$$J_\alpha(x) = \frac{1}{2\pi i}\int_{-\infty}^{(0^+)} t^{-\alpha-1}\exp\left\{\frac{x}{2}\left(t - \frac{1}{t}\right)\right\} dt$$

   as $x \to +\infty$, where $\alpha$ is any complex number, and the loop contour starts from $-\infty$ on the real axis, passes around the origin in the counterclockwise direction, and returns to $-\infty$.
9. Consider the function

$$F(z) = \int_0^\infty \exp\left\{iz\left(\frac{t^3}{3} + t\right)\right\} dt, \qquad 0 < \arg z < \pi.$$

Show that, by rotating the path of integration through an angle $\psi$, $0 < \psi < \frac{1}{3}\pi$, $F(z)$ can be continued analytically to the region $-\frac{1}{3}\pi < \arg z < \pi$. Use the steepest descent method to derive the asymptotic expansion

$$F(z) \sim i \sum_{m=0}^{\infty} \frac{(3m)!}{m!3^m} z^{-2m-1} \tag{22.4.5}$$

as $z \to \infty$, uniformly in the sector $-\frac{1}{3}\pi + \delta \le \arg z \le \pi - \delta$, $\delta > 0$. Note that this expansion holds, in particular, as $z \to +\infty$ through positive real values. Also, since $F(-z) = \overline{F(z)}$, the domain of validity of (22.4.5) can be extended to $-\frac{1}{3}\pi + \delta \le \arg z \le \pi + \frac{1}{3}\pi - \delta$.

## Remarks and further reading

For more on integral methods in asymptotic analysis, see Wong [145].

The functional equation (22.3.6) is obviously equivalent to the functional equation for the Dedekind eta function (22.4.3):

$$\eta\left(-\frac{1}{z}\right) = \sqrt{\frac{z}{i}}\,\eta(z). \tag{22.4.6}$$

The eta function is closely related to the discriminant $\Delta = g_2^3 - 27g_3^2$ of the polynomial associated to the Weierstrass $\wp$ function; see Section 17.4. In fact

$$\Delta(z) = (2\pi)^{12}\eta^{24}(z).$$

For this and role of the eta function in number theory, see Apostol [9], Chapter 3.

# Chapter 23
# Complex interpolation and the Riesz–Thorin theorem



Often a given vector space has more than one natural norm. For example, the space of bounded continuous functions $f : [0,1] \to \mathbb{C}$ can be equipped with the norms

$$||f||_1 = \int_0^1 |f(x)|\,dx, \qquad ||f||_\infty = \sup_{x\in[0,1]} |f(x)|.$$

There are several natural ways to *interpolate* between these, i.e. to find a scale of norms $||f||_p$, $1 < p < \infty$ that, in some suitable sense, have $||f||_1$ and $||f||_\infty$ as limits. One desirable property is that if $T$ is a map from the space to itself that is continuous with respect to the extreme norms,

$$||Tf||_1 \le C_1||f||_1, \qquad ||Tf||_\infty \le C_\infty||f||_\infty, \tag{23.0.1}$$

then $T$ is also continuous with respect to the intermediate norms:

$$||Tf||_p \le C_p||f||_p, \qquad 1 < p < \infty. \tag{23.0.2}$$

Complex interpolation is a very general method for constructing a scale of intermediate spaces between two normed linear spaces, in such a way that the passage from (23.0.1) to (23.0.2) is valid.

The Riesz–Thorin theorem is a generalization of the passage from (23.0.1) to (23.0.2) in the context of $L^p$ spaces. This theorem preceded the method of complex interpolation. Now, as here, it is often presented as a consequence of the method.

## 23.1 Interpolation: the complex method

A *norm* on a real or complex vector space $X$ is a non-negative function $u \to ||u||$, $u \in X$, with the properties:



R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_23

$$||u|| = 0 \quad \text{if and only if} \quad u = 0;$$
$$||au|| = |a|\,||u|| \quad \text{if} \quad a \text{ is a scalar;}$$
$$||u+w|| \le ||u|| + ||w||, \quad \text{for every pair } u, w \text{ in } X.$$

A norm induces a metric $d(u,w) = ||u-w||$. A *normed linear space* is a vector space equipped with a norm. A *Banach space* is a normed linear space that is complete with respect to the induced metric.

Suppose that $X_0$ and $X_1$ are two Banach spaces, with norms $||\ ||_0$, $||\ ||_1$. Suppose also that the intersection $X_0 \cap X_1$ is dense in each space. Thus $X_j$ can be thought of as the completion of $X_0 \cap X_1$ with respect to the norm $||\ ||_j$, $j = 0, 1$. In this context, the basic problem is to find a natural way to construct intermediate spaces $X_\theta$, $0 < \theta < 1$ with norms $||\ ||_\theta$.

A linear map $T : X \to Y$ from one normed linear space to another is said to be *bounded* if there is a constant $C$ such that $||Tu||_Y \le C||u||_X$, all $u \in X$. Here, of course, $||\ ||_X$, $||\ ||_Y$ denote the norms in $X$ and $Y$, respectively. If $T$ is bounded, the smallest such constant is denoted by $||T||$. Clearly

$$||T|| = \sup_{||u||_X = 1} ||Tu||_Y.$$

If one has an interpolation method, a natural question is the following. Suppose that we interpolate between spaces $X_0$, $X_1$ and also between spaces $Y_0$ and $Y_1$. By an abuse of notation, we use $||\ ||_\theta$ to denote the norm in $X_\theta$ or in $Y_\theta$, according to the context. Suppose now that $T$ is a linear map,

$$T : X_0 \cap X_1 \to Y_0 \cap Y_1,$$

and $T$ is bounded with respect to the norms $||\ ||_0$ and $||\ ||_1$:

$$||Tu||_0 \le C_0 ||u||_0, \qquad ||Tu||_1 \le C_1 ||u||_1.$$

Then is $T$ also bounded as a map from $X_\theta$ to $Y_\theta$, $0 < \theta < 1$?

The complex method of interpolation is based on Hadamard's *three lines theorem.*

**Theorem 23.1.1.** *Suppose that the function $f$ is bounded and holomorphic in the vertical strip $\{z : 0 < \operatorname{Re} z < 1\}$ and continuous on the closure. Let*

$$M_\theta = \sup_{\operatorname{Re} z = \theta} |f(z)|.$$

*Assume $M_0 M_1 > 0$. Then*

$$M_\theta \le M_0^{1-\theta} M_1^\theta. \tag{23.1.1}$$

*Proof:* The first step is to prove that for each $z$ in the strip,

$$|f(z)| \le \max\{M_0, M_1\}. \tag{23.1.2}$$

Given $\varepsilon > 0$, let $f_\varepsilon(z) = e^{\varepsilon(z^2-1)} f(z)$. Since $\mathrm{Re}\,[(\theta+it)^2-1] \le -t^2$, it follows that

$$|f_\varepsilon(\theta+it)| \le e^{-\varepsilon t^2}|f(\theta+it)|, \quad 0\le\theta\le 1, \;\; t\in\mathbb{R}.$$

Since $f$ is assumed to be bounded in the strip, it follows that for sufficiently large $t^2$,

$$|f_\varepsilon(\theta+it)| \le \max\{M_0, M_1\}. \tag{23.1.3}$$

By the maximum principle, this inequality holds throughout the truncated strip

$$S_R = \{z : 0 \le \mathrm{Re}\,z \le 1, \; |\mathrm{Im}\,z| \le R\}$$

for $R$ sufficiently large. Therefore the inequality (23.1.3) holds throughout the strip. Since $\mathrm{Re}\,[1-(\theta+it)^2] \le 1+t^2$, it follows that

$$|f(\theta+it)| = |e^{\varepsilon(1-z^2)} f_\varepsilon(\theta+it)| \le e^{\varepsilon(1+t^2)} \max\{M_0, M_1\}$$

throughout the strip. Taking $\varepsilon \to 0$, we obtain (23.1.2).

To obtain the estimate (23.1.1), let

$$g(z) = M_0^{z-1} M_1^{-z} f(z).$$

Then $g$ is bounded and holomorphic in the strip, and $|g(z)| \le 1$ on the boundary of the strip. Therefore $|g| \le 1$ in the strip and

$$|f(\theta+it)| = M_0^{1-\theta} M_1^{\theta} |g(\theta+it)| \le M_0^{1-\theta} M_1^{\theta}. \qquad \square$$

Suppose now that we want to interpolate between spaces $X_0$ and $X_1$, as above. The general procedure is to consider the family $\mathcal{F}$ of functions $f$ that have the following properties:

(i) $f$ is a continuous map from the strip with values in the vector space sum $X_0 + X_1 = \{u_0 + u_1; u_j \in X_j\}$,

(ii) $f$ is holomorphic in the interior of the strip,

(iii) $f(it)$ belongs to $X_0$, $f(1+it)$ belongs to $X_1$, and

$$(||f|| \equiv \max\left\{\sup_{t\in\mathbb{R}} ||f(it)||_0, \sup_{t\in\mathbb{R}} ||f(1+it)||_1\right\} < \infty.$$

Then for $0 < \theta < 1$, $X_\theta$ consists of those elements $u \in X_0 + X_1$ such that $f(\theta) = u$ for some $f \in \mathcal{F}$. The norm is defined by

$$||u||_{[\theta]} = \inf_{f\in\mathcal{F}, f(\theta)=u} ||f||. \tag{23.1.4}$$

Here we have been vague about the terms "continuous" and "holomorphic" in the general case. They will be clear in the cases to be considered here.

## 23.2 $L^p$ spaces

The model cases we consider here come under the heading of "$L^p$ spaces," $1 \leq p < \infty$. They are, informally,

**(a)** $L^p(\mathbb{R})$, the space of functions $u : \mathbb{R} \to \mathbb{C}$ such that

$$||u||_p = \left[\int_{\mathbb{R}} |u(x)|^p \, dx\right]^{1/p} < \infty, \qquad 1 \leq p < \infty.$$

For $p = \infty$ the norm is taken to be the (essential) supremum of $|u(x)|$.

**(b)** The periodic versions $L^p_{\text{per}}(\mathbb{R})$ of the preceding spaces. Here the line $\mathbb{R}$ is replaced by the interval $(-\pi, \pi)$ and the norm is

$$||u||_p = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} |u(\theta)|^p \, d\theta\right]^{1/p}, \qquad 1 \leq p < \infty$$

with a corresponding change for $p = \infty$.

**(c)** $l^p(\mathbb{Z})$, the space of two-sided complex sequences $\mathbf{a} = (a_n)_{n=-\infty}^{\infty}$, with

$$||\mathbf{a}||_p = \left[\sum_{n=-\infty}^{\infty} |a_n|^p\right]^{1/p}, \qquad 1 \leq p < \infty$$

and $||\mathbf{a}||_\infty = \sup |a_n|$. In the following, we will often write the sum here in the form of an integral. (It is in fact an integral with respect to the "counting measure".)

A detailed study of examples (a) and (b) requires measure theory. Here we shall work within the space $X$, consisting of continuous functions $f$ that vanish outside a bounded interval (depending on $f$) in case (a) or are periodic, in case (b). The $L^p$ spaces (for $1 \leq p < \infty$) are the completions of $X$ with respect to the corresponding metrics. (In case (c), $X$ consists of sequences each of which has only finitely many non-zero terms.) Some adjustments need to be made when $p = \infty$.

The arguments carry over to very general $L^p$ spaces. Here we use results from Chapter 1.

**Lemma 23.2.1.** *In each of the cases (a), (b), or (c) above, one has Hölder's inequality; in the notation of case (a) it is*

$$\left|\int_{\mathbb{R}} u(x)w(x) \, dx\right| \leq ||u||_p ||w||_q \quad \text{if} \quad \frac{1}{p} + \frac{1}{q} = 1. \tag{23.2.1}$$

*Conversely, for each $u \in L^p$,*

$$||u||_p = \sup_{w \in L^q, ||w||_q = 1} \int_{\mathbb{R}} |u(x)w(x)| \, dx. \tag{23.2.2}$$

It seems that any reasonable method of interpolation between two $L^p$ spaces of the same type should produce the $L^p$ spaces with intermediate values of $p$. This is indeed the case here.

**Theorem 23.2.2.** *In each of the cases (a), (b), or (c), let $X_0 = L^{p_0}$ and $X_1 = L^{p_1}$, $p_0 \neq p_1$. Then for $0 < \theta < 1$, the interpolation space $X_\theta$ is $L^{p_\theta}$, where*

$$\frac{1}{p_\theta} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}.$$

*Proof:* We use the notation associated with case (a). Given $u$ in the space $X$ of continuous functions with bounded support, and $0 < \theta < 1$, we may normalize with

$$\int_{\mathbb{R}} |u(x)|^{p_\theta}\, dx = 1, \qquad \frac{1}{p_\theta} = (1-\theta)\frac{1}{p_0} + \theta\frac{1}{p_1}. \tag{23.2.3}$$

Define $f$ as a map from the strip $\{z : 0 \le \operatorname{Re} z \le 1\}$ to $X$ by

$$f(z,x) = \begin{cases} 0 & \text{if } u(x) = 0; \\ |u(x)|^{\alpha(z)} \dfrac{u(x)}{|u(x)|} & \text{if } u(x) \neq 0, \end{cases}$$

where

$$\alpha(z) = (1-z)\frac{p_\theta}{p_0} + z\frac{p_\theta}{p_1}. \tag{23.2.4}$$

For a given $x$, $|f(z,x)|$ is independent of $\operatorname{Im} z$. Equations (23.2.3), (23.2.4) imply that

$$||f(it,\cdot)||_{p_0} = 1 = ||f(1+it,\cdot)||_{p_1}.$$

so the norm $||u||_{[\theta]}$ of $u$ as an element of the interpolation space $X_\theta$ is at most 1, which is $||u||_{p_\theta}$.

Suppose now that $w$ is in $L^{q_\theta}$, where $1/p_\theta + 1/q_\theta = 1$, and suppose also that $||w||_{q_\theta} = 1$. Choose indices $q_0$ and $q_1$ with

$$\frac{1}{p_0} + \frac{1}{q_0} = 1 = \frac{1}{p_1} + \frac{1}{q_1},$$

and let $Y_0 = L^{q_0}$, $Y_1 = L^{q_1}$. The index $q_\theta$ above satisfies

$$\begin{aligned} \frac{1}{q_\theta} = 1 - \frac{1}{p_\theta} &= 1 - \left[(1-\theta)\frac{1}{p_0} + \theta\frac{1}{p_1}\right] \\ &= (1-\theta)\left[1 - \frac{1}{p_0}\right] + \theta\left[1 - \frac{1}{p_1}\right] \\ &= (1-\theta)\frac{1}{q_0} + \theta\frac{1}{q_1}. \end{aligned}$$

By the argument just given for $u$, the interpolation norm $||w||_{[\theta]}$ is at most 1. Given $\varepsilon > 0$, let $g$ and $h$ be two elements of $\mathcal{F}$ such that $g(\theta) = u$ and $h(\theta) = w$ and

$||g|| \le ||u||_{[\theta]} + \varepsilon$, $||h|| \le ||w||_{[\theta]} + \varepsilon$. Then by Hölder's inequality

$$\begin{aligned}\left|\int_{\mathbb{R}} u(x)w(x)\,dx\right| &= \left|\int_{\mathbb{R}} g(\theta,x)h(\theta,x)\,dx\right| \\ &\le (||u||_{[\theta]} + \varepsilon)(||w||_{[\theta]} + \varepsilon) \le (||u||_{[\theta]} + \varepsilon)(1+\varepsilon).\end{aligned}$$

In view of the second part of Lemma 23.2.1, this implies that $||u||_{p_\theta} = ||u||_{[\theta]}$. □

## 23.3 Application: the Riesz–Thorin theorem

**Theorem 23.3.1.** (Riesz–Thorin) *Suppose that $X_0 = L^{p_0}$ and $X_1 = L^{p_1}$ are $L^p$ spaces of the same type (i.e. with respect to the same measure), and suppose that $Y_0 = L^{q_0}$ and $Y_1 = L^{q_1}$ are of the same type as each other, but not necessarily of the same type as the $X_j$. Suppose that $T$, defined on $X_0 \cap X_1$, extends as a bounded map from $X_0$ to $Y_0$ with norm $||T||_0$ and as a bounded map from $X_1$ to $Y_1$ with norm $||T||_1$. Then for $0 < \theta < 1$, $T$ extends as a bounded map from the interpolation space $X_\theta$ to $Y_\theta$, with norm*

$$||T||_\theta \le ||T||_0^{1-\theta}||T||_1^\theta. \tag{23.3.1}$$

*Proof:* If $p_0 = p_1$ and $q_0 = q_1$, there is nothing to prove. Suppose that $u$ and $w$ are bounded, continuous functions with bounded support. Given $0 < \theta < 1$, we normalize with

$$||u||_{p_\theta} = 1 = ||w||_{q'_\theta},$$

where

$$\begin{aligned}\frac{1}{p_\theta} &= (1-\theta)\frac{1}{p_0} + \theta\frac{1}{p_1}; \\ \frac{1}{q'_\theta} &= (1-\theta)\frac{1}{q'_0} + \theta\frac{1}{q'_1} = (1-\theta)\left(1-\frac{1}{q_0}\right) + \theta\left(1-\frac{1}{q_1}\right).\end{aligned}$$

Let $f(z,x) = 0$ where $u(x) = 0$, $g(z,x) = 0$ where $w(x) = 0$, and otherwise

$$f(z,x) = |u(x)|^{\alpha(z)}\frac{u(x)}{|u(x)|}, \qquad g(z,x) = |w(x)|^{\beta(z)}\frac{w(x)}{|w(x)|},$$

where

$$\alpha(z) = (1-z)\frac{p_\theta}{p_0} + z\frac{p_\theta}{p_1}, \qquad \beta(z) = (1-z)\frac{q'_\theta}{q'_0} + z\frac{q'_\theta}{q'_1}.$$

Then

$$\begin{aligned}||f(it,\cdot)||_{p_0} &= ||f(1+it,\cdot)||_{p_1} = ||u||_{p_\theta} = 1, \\ ||g(it,\cdot)||_{q'_0} &= ||g(1+it,\cdot)||_{q'_1} = ||w||_{q'_\theta} = 1,\end{aligned}$$

so

$$\int_{\mathbb{R}} |Tf(it,x)\, g(it,x)|\, dx \le ||T||_0;$$

$$\int_{\mathbb{R}} |Tf(1+it,x)\, g(1+it,x)|\, dx \le ||T||_1.$$

Consequently

$$\left| \int_{\mathbb{R}} Tu(x)\, w(x)\, dx \right| \le ||T||_0^{1-\theta} ||T||_1^{\theta}.$$

It follows that $||T||_\theta \le ||T||_0^{1-\theta} ||T||_1^{\theta}$. □

## 23.4 Application to Fourier series

We refer here to Chapter 1 and Chapter 4. If $u : \mathbb{R} \to \mathbb{C}$ is continuous and periodic, with period $2\pi$, then the Fourier coefficients $\widehat{u}$ are defined by

$$\widehat{u}(k) \;=\; \frac{1}{2\pi i} \int_{-\pi}^{\pi} u(x)\, e^{-ikx}\, dx, \qquad k = 0, \pm 1, \pm 2 \ldots .$$

Since $|e^{-ikx}| \equiv 1$, it is clear that the map

$$Tu \;=\; \{\widehat{u}(k)\}_{k=-\infty}^{\infty}$$

extends to map $L^1_{\text{per}}(\mathbb{R})$ to $l^\infty(\mathbb{Z})$ and

$$||Tu||_\infty \;\le\; ||u||_1;$$

equality is attained if $u \ge 0$.

The Riesz–Fischer theorem, Theorem 4.4.3, says that $T$ extends to map $L^2_{\text{per}}(\mathbb{R})$ to $l^2(\mathbb{Z})$, and

$$||Tu||_2 \;=\; ||u||_2.$$

Therefore $T$ extends to map $L^p_{\text{per}}(\mathbb{R})$ to $l^q(\mathbb{Z})$ and

$$||Tu||_q \;\le\; ||u||_p, \quad \text{for} \quad 1 \le p \le 2, \quad \frac{1}{p} + \frac{1}{q} = 1. \tag{23.4.1}$$

Similarly the dual map from sequences to functions,

$$\mathbf{a} \;=\; \{a_k\}_{-\infty}^{\infty} \;\to\; T^*\mathbf{a} \;=\; \sum_{-\infty}^{\infty} a_k\, e^{ikx} \tag{23.4.2}$$

takes $l^1(\mathbb{Z})$ to the space of bounded continuous periodic functions and maps $l^2(\mathbb{Z})$ onto $L^2_{\text{per}}(\mathbb{R})$, each with norm 1, so

$$||T^*\mathbf{a}||_q \;\le\; ||\mathbf{a}||_p, \quad \text{for} \quad 1 \le p \le 2, \quad \frac{1}{p}+\frac{1}{q}=1. \tag{23.4.3}$$

The inequalities (23.4.1) and (23.4.3) are the *Hausdorff–Young inequalities*.

## Exercises

1. Prove the following consistency result for iterated interpolation. Let $Y_0 = X_{\theta_0}$ and $Y_1 = X_{\theta_1}$, where $0 \le \theta_0 < \theta_1 \le 1$ and the $X_{\theta_j}$ are interpolations spaces for $X_0$ and $X_1$. Then for $0 < \alpha < 1$, $Y_\alpha = X_\theta$ with $\theta = (1-\alpha)\theta_0 + \alpha\theta_1$.
2. Suppose that $w_0$ and $w_1$ are positive continuous functions on $\mathbb{R}$. Let

$$X_j \;=\; L^p(\mathbb{R}, w_j dx) \;=\; \{f : \mathbb{R} \to \mathbb{C} : \int_{-\infty}^{\infty} |f(x)|\, w_j(x)\, dx < \infty\}, \quad j = 0, 1,$$

where $1 \le p < \infty$. Prove that the corresponding interpolation space $X_\theta$ is $L^p(\mathbb{R}, w_\theta dx)$, where

$$w_\theta(x) \;=\; w_0(x)^{1-\theta} w_1(x)^\theta, \quad 0 < \theta < 1.$$

3. The spaces $H^s(\mathbb{R})$ (and the corresponding spaces in higher dimensions $H^s(\mathbb{R}^n)$) are useful in the study of differential equations. For $m$ a non-negative integer, $H^m(\mathbb{R})$ can be defined as the completion of the space of Schwartz functions with respect to the norm

$$||u||^2_{H^m} \;=\; \sum_{j=0}^{m} \binom{m}{j} ||u^{(j)}||^2.$$

The spaces $H^s(\mathbb{R})$, $0 < s < m$, are defined by interpolation. Describe these spaces explicitly, and show that for integers $0 < k < m$ the two definitions of $H^k(\mathbb{R})$ are equivalent. (Use the Fourier transform and Exercise 2.)
4. Let $A = (a_{ij})$ be an infinite matrix with

$$\sup_{i,j \in \mathbb{Z}} |a_{ij}| \;=\; M < \infty. \tag{23.4.4}$$

(a) Prove that the map $\mathbf{x} \to A\mathbf{x}$,

$$(A\mathbf{x})_i \;=\; \sum_{j=-\infty}^{\infty} a_{ij} x_j$$

maps $L^p(\mathbb{Z})$ to itself with norm $M$.

(b) Suppose that $A$ is unitary:

$$\sum_{j=-\infty}^{\infty} a_{ij} \bar{a}_{jk} \;=\; \begin{cases} 1 \text{ if } i = k, \\ 0 \text{ if } i \ne k. \end{cases}$$

Prove that $A$ maps $L^2(\mathbb{Z})$ to itself with norm 1.

(c) Assuming both (23.4.4) and that $A$ is unitary, prove that $A$ maps $L^p(Z)$ to $L^{p'}$, $1 < p < 2$, $1/p + 1/p' = 1$, with norm at most $M^{1/p-1/p'}$.

5. Prove Stein's generalization [130] of the Riesz–Thorin theorem: under the other assumptions of that theorem, assume that the map $T = T(z)$ depends on $z$, that for each pair of piecewise constant functions that are integrable with respect to the measure $\mu$ associated to the $X_j$ and with respect to the measure $\nu$ associated to $Y_j$, respectively, the map

$$z \to \int [T(z)f](y)\, g(y)\, d\mu(y)$$

is holomorphic in the strip $0 < \operatorname{Re} z < 1$, and continuous on the closure of the strip. Suppose also that $T(z)$ is uniformly bounded as a map from $X_0$ to $Y_0$ for $\operatorname{Re} z = 0$ and from $X_1$ to $Y_1$ for $\operatorname{Re} z = 1$. Then $T(\theta)$ extends to a bounded map from $X_\theta$ to $Y_\theta$.

## Remarks and further reading

Interpolation theory plays a large role in real analysis, functional analysis, and approximation theory. The complex method was developed circa 1960 by Calderón [30] and Lions [92]. There are other methods of interpolation based on real analysis. Both the complex method and the real methods are treated extensively in Bergh and Löfström [18].

# References


1. Abikoff, W.: The uniformization theorem. Amer. Math. Monthly **88**, 574–592 (1981)
2. Ablowitz, M.J., Fokas, A.S.: Complex Variables. Introduction and Applications, 2nd edn. Cambridge University Press, Cambridge (2003)
3. Agmon, S.: Asymptotic formulas with remainder estimates for eigenvalues of elliptic operators. Arch. Rat. Mech. Anal. **28**, 165–183 (1967)
4. Ahlfors, L.: Über eine Methode in der Theorie der meromorphen Funktionen. Soc. Sci. Fenn. Comm. Phys.-Math.**8**, no. 10 (1935)
5. Ahlfors, L.: Complex Analysis, 3rd edn. McGraw-Hill, New York (1978)
6. Ahlfors, L., Bers, L.: Riemann's mapping theorem for variable metrics. Ann. Math. **72**, 385–404 (1960)
7. Andrews, G.E., Askey, R., Roy, R.: Special Functions. Cambridge University Press, Cambridge (1999)
8. Apostol, T.M.: Introduction to Analytic Number Theory. Springer, New York, Heidelberg, Berlin (1976)
9. Apostol, T.M.: Modular Functions and Dirichlet Series in Number Theory, 2nd edn. Springer, New York, Heidelberg, Berlin (1990)
10. Armitage, J.V., Eberlein, W.F.: Elliptic Functions. Cambridge University Press, Cambridge (2006)
11. Arthur, J.: *L*-functions and automorphic representations. Proc. Intnl. Congr. Math. Seoul **1**, 171–197 (2014)
12. Artin, E.: The Gamma Function. Holt. Rinehart, and Winston, New York, 1984, Dover, New York (2015)
13. Astala, K., Iwaniec, K., Martin, G.: Elliptic Partial Differential Equations and Quasiconformal Mappings in the Plane. Princeton University Press, Princeton (2009)
14. Axler, S., Bourdon, P., Ramey, W.: Harmonic Function Theory, 2nd edn. Springer, New York, Heidelberg, Berlin (2001)
15. Beals, R., Szmigielski, J.: Meijer *G*-functions, a gentle introduction. Notices Amer. Math. Soc. **60**, 866–872 (2013)




R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8_0

16. Beals, R., Wong, R.: Special Functions and Orthogonal Polynomials. Cambridge University Press, Cambridge (2016)
17. Bell, S.R.: The Cauchy Transform, Potential Theory, and Conformal Mapping, 2nd edn. Chapman and Hall/CRC, Boca Raton, FL (2016)
18. Bergh, J., Löfström, J.: Interpolation Spaces. An Introduction. Springer, New York, Heidelberg, Berlin (1976)
19. Bergweiler, W.: An Introduction to Complex Dynamics. University Coimbra, Coimbra (1995)
20. Berndt, B.C.: A new proof of the functional equation for Dirichlet $L$-functions. Proc. Amer. Math. Soc. **37**, 355–357 (1973)
21. Bers, L.: Uniformization by Beltrami equations. Comm. Pure Appl. Math. **14**, 215–228 (1961)
22. Bers, L.: Quasiconformal mappings, with applications to differential equations, function theory, and topology. Bull. Amer. Math. Soc. **83**, 1083–1100 (1977)
23. Bieberbach, L.: Über die Koeffizienten derjenigen Potenzreihen, welche eine schlichte Abbildung des Einheitskreise vermitteln, pp. 940–955. S.-B. Preuss. Akad, Wiss (1916)
24. Boas, R.: Entire Functions. Academic Press, New York (1954)
25. Bohr, H., Mollerup, J.: Laerebog i Matematisk Analyse, vol. 3. J. Gjellerup, Copenhagen (1922)
26. Bolza, O.: On binary sextics with linear transformations into themselves. Amer. J. Math. **10**, 47–70 (1887)
27. Borwein, P., Choi, S., Rooney, B., Weirathmueller, A. (eds.): The Riemann Hypothesis. Springer, New York, Heidelberg, Berlin (2008)
28. Brezhnev, Yu.V.: On uniformization of Burnside's curve $y^2 = x^5 - x$. J. Math. Phys.**50**, no. 10, 103519, 23 pp (2009)
29. Burnside, W.S.: Note on the equation $y^2 = x(x^4 - 1)$. Proc. London Math. Soc. **24**, 17–20 (1893)
30. Calderón, A.P.: Intermediate spaces and interpolation, the complex method. Studia Math. **24**, 113–190 (1964)
31. Cavalieri, R.: Riemann Surfaces and Algebraic Curves. Cambridge University Press, Cambridge (2016)
32. Cherry, W., Ye, Zh: Nevanlinna's Theory of Value Distribution. Springer, New York, Heidelberg, Berlin (2001)
33. Choimet, D., Queféllec, H.: Twelve Landmarks of Twentieth Century Analysis. Cambridge University Press, Cambridge (2015)
34. de Branges, L.: A proof of the Bieberbach conjecture. Acta Mathematica **154**, 137–152 (1985)
35. de la Vallée Poussin, C.-J.: Recherches analytiques sur la théorie des nombres premiers. Ann. Soc. Scient. Bruxelles **20**, 183–256 (1896)
36. Donaldson, S.: Riemann Surfaces. Oxford University Press, Oxford (2011)
37. Dou, Z.-L., Zhang, Q.: Six Short Chapters on automorphic forms and $L$-Functions. Springer, New York, Heidelberg, Berlin (2012)

38. Dubrovin, B.A.: Theta functions and nonlinear equations. Russian Math. Surveys **36**(2), 11–92 (1981)
39. Duncan, J.F.R., Griffin, M.J., Ono, K.: Moonshine. arXiv:1411.6571
40. Duren, P.: Univalent Functions. Springer, New York, Heidelberg, Berlin (1983)
41. Dutka, J.: The early history of the factorial function. Arch. Hist. Exact Sci. **43**, 225–249 (1991/1992)
42. Edwards, H.M.: Riemann's Zeta Function. Dover, Mineola N. Y (1974)
43. Estrada, R.: Singular Integral Equations. Birkhäuser, Boston (2000)
44. Farkas, H.M., Kra, I.: Riemann Surfaces, 2nd edn. Springer, New York, Heidelberg, Berlin (1992)
45. Farkas, H.M., Kra, I.: Theta Constants, Riemann Surfaces, and the Modular Group. Springer, New York, Heidelberg, Berlin (2001)
46. Fatou, P.: Séries trigonometriques et séries de Taylor. Acta Math. **30**, 335–400 (1906)
47. Gårding, L.: Some Points of Analysis and Their History. Amer. Math. Soc, Providence (1997)
48. Gelbart, S.: An elementary introduction to the Langlands program. Bull. Amer. Math. Soc. (N.S.).**10**, 177–219 (1984)
49. Gohberg, I.C., Krein, M.R.: Systems of integral equations on a half line with kernel depending on the difference of arguments. Uspehi Mat. Nauk**13**, no. 2, 3–72 (1958), Amer. Math. Soc. Translations, ser.2, vol. 14, 217–288 (1960)
50. Goldberg, A., Ostrovskii, I.: Value Distribution of Meromorphic functions. Amer. Math. Soc. Translations, vol. 236, Amer. Math. Soc., Providence (2008)
51. Goldschmidt, D.M.: Algebraic Functions and Projective Curves. Springer, New York, Heidelberg, Berlin (2003)
52. Grafakos, L.: Modern Fourier Analysis, 3rd edn. Springer, New York, Heidelberg, Berlin (2014)
53. Gray, J.: On the history of the Riemann mapping theorem. Rend. Circ. Mat. Palermo. **34**, 47–94 (1994)
54. Gronwall, T.H.: Some remarks on conformal representation. Annals Math. **16**, 72–76 (1914–1915)
55. Hadamard, J.: Étude sur les propriétés des fonctions entières et en particulier d'une fonction considérée par Riemann. J. Math. Pures Appl. [4] **9**, 171–216 (1893)
56. Hadamard, J.: Sur la distributions des zéros de la fonction $\zeta(s)$ et ses conséquences arithmétiques. Bull. Soc. Math. France **24**, 199–220 (1896)
57. Hardy, G.H.: A theorem concerning Fourier transforms. J. London Math. Soc. **8**, 227–231 (1933)
58. Hardy, G.H.: Notes on special systems of orthogonal functions. IV. The orthogonal functions of Whittaker's cardinal series. Proc. Cambridge Philos. Soc. **37**, 331–348 (1941)
59. Hardy, G.H.: Divergent Series. Clarendon Press, Oxford (1949)

60. Hardy, G.H., Littlewood, J.E.: Tauberian theorems concerning power series and and Dirichlet's series whose terms are positive. Proc. London Math. Soc. ser. **2**(13), 174–191 (1914)
61. Hayman, W.K.: Meromorphic Functions. Oxford University Press, London (1964)
62. Hayman, W.K.: Subharmonic Functions, vol. 2. Academic Press, London (1989)
63. Hayman, W.K., Kennedy, P.B.: Subharmonic Functions, vol. 1. Academic Press, London, New York (1976)
64. Hille, E.: Analytic Function Theory, vol. I, II. Ginn & Co., Boston, 1959 (1962)
65. Hille, E.: Ordinary Differential Equations in the Complex Domain. Dover, Mineola N. Y (1997)
66. Hörmander, L.: An Introduction to Complex Analysis in Several Variables, 3rd edn. North-Holland, Amsterdam (1990)
67. Hu, P.-C., Yang, C.-C.: Value Distribution Theory Related to Number Theory. Birkhäuser, Basil (2006)
68. Hubbard, J.H.: Teichmüller Theory and Applications to Geometry, Topology, and Dynamics, vol. 1. Matrix Editions, Ithaca (2006)
69. Ikehara, S.: An extension of Landau's theorem in the analytic theory of numbers. J. Math. Phys. M.I.T. **10**, 1–12 (1931)
70. Ireland, K., Rosen, M.: A Classical Introduction to Modern Number Theory, 2nd edn. Springer, New York, Heidelberg, Berlin (1990)
71. Iwaniec, H., Sarnak, P.: Perspectives on analytic $L$-functions. Geom. Anal. Funct. Anal. Special Vol. pp. 705–741 (2000)
72. Ivic, A.: The Riemann Zeta Function. Dover, Mineloa, N.Y. (1985)
73. Jensen, J.L.V.W.: Sur un nouvel et important théorème de la théorie des fonctions. Acta Math. **22**, 359–364 (1899)
74. Kawakubo, K.: The Theory of Transformation Groups. Oxford University Press, Oxford (1991)
75. Karamata, J.: Über die Hardy-Littlewoodschen Umkehrungen des Abelschen Stetigkeitsatzes. Math. Z. **32**, 319–320 (1930)
76. Kempf, G.: Complex Abelian Varieties and Theta Functions. Springer, New York, Heidelberg, Berlin (1991)
77. Korevaar, J.: Tauberian Theory, a Century of Development. Springer, New York, Heidelberg, Berlin (2004)
78. von Koch, H.: Sur la distribution des nombres premiers. Acta Mathematica **24**, 159–182 (1901)
79. Krantz, S.G.: Harmonic and Complex Analysis in Several Variables. Springer, Cham (2017)
80. Krein, M.R.: Integral equations on a half line with kernel depending on the difference of arguments. Uspehi Mat. Nauk.**13**, no. 5, 3–120 (1958) Amer. Math, Soc. Translations, ser. 2, vol. 22, 163–288 (1962)
81. Kuusalo, T., Näätänen, M.: Geometric uniformization in genus 2. Ann. Acad. Sc. Fenn. **20**, 401–418 (1995)

82. Kythe, P.K.: Handbook of Conformal Mappings and Applications. CRC Press, Boca Raton (2019)
83. Langlands, R.: *L*-functions and automorphic representations. Proc. Intnl. Cong. Math. Helsinki, pp. 165–175 (1978)
84. Lawrie, J.B., Abrahams, I.D.: A brief historical perspective of the Wiener-Hopf technique. J. Engrg. Math. **59**, 351–358 (2007)
85. Lehner, J.: Discontinuous Groups and Automorphic Functions. Amer. Math. Soc, Providence (1964)
86. Lehner, J.: A Short Course in Automorphic Functions. Holt, Rinehart and Winston, New York (1966)
87. Lehto, O.: Univalent Functions and Teichmüller Spaces. Springer, New York, Heidelberg, Berlin (1987)
88. Levin, B.Ya.: Distribution of zeros of entire functions, revised edn. Amer. Math. Soc., Providence (1980)
89. Levin, B.Ya.: Lectures on Entire Functions. Amer. Math. Soc., Providence (1996)
90. Lévy, P.: Sur la convergence absolue des séries de Fourier. Comp. Math. **1**, 1–14 (1935)
91. Ling, S., Wang, H., Xing, C.: Algebraic Curves and Cryptography. CRC Press, Boca Raton (2018)
92. Lions, J.-L.: Une construction d'espaces d'interpolation. C. R. Acad. Sci. Paris **251**, 1853–1855 (1960)
93. Littlewood, J.E.: The converse of Abel's theorem on power series. Proc. London Math. Soc. ser. **2**(9), 443–448 (1911)
94. Àlvaro, Lozano-Robledo: Curves, Elliptic, Forms, Modular: and Their *L*-functions. Amer. Math. Soc, Providence (2011)
95. Malliavin, P.: Un théorème taubérien avec reste pour la transformée de Stieltjes. C. R. Acad. Sci. Paris. **255**, 2351–2352 (1962)
96. von Mangoldt, H.: Zu Riemanns Abhandlung "Ueber die Anzahl der Primzahlen unter einen gegebenen Grösse". J. Reine Angew. Math. **114**, 255–305 (1895)
97. Marden, A.: Hyperbolic Manifolds. An Introduction in 2 and 3 Dimensions. Cambridge University Press, Cambridge (2016)
98. Milne, E.A.: Radiative Equilibrium in the outer layer of a star. Monthly Notices Roy. Acad. Sci. **81**, 361–375 (1921)
99. Miranda, R.: Algebraic Curves and Riemann Surfaces. Amer. Math. Soc, Providence (1995)
100. Moreno, C.J.: Advanced Analytic Number Theory: *L*-Functions. Amer. Math. Soc, Providence (2005)
101. Murty, M.R. (ed.): Theta Functions: From the Classical to the Modern. Amer. Math. Soc, Providence (1993)
102. Narasimhan, R.: Compact Riemann Surfaces. Birkhäuser, Basel (1992)
103. Nehari, Z.: The Schwarzian derivative and schlicht functions. Bull. Amer. Math. Soc. **55**, 545–551 (1949)

104. Nehari, Z.: Conformal Mapping. McGraw Hill, London (1952), Dover, New York (1975)
105. Nevanlinna, R.: Le Théorème de Picard-Borel et la Théorie des Fonctions Méromorphes. Gauthier-Villars, Paris (1929) Chelsea, New York (1974)
106. Newman, D.J.: A simple proof of Wiener's $1/f$ theorem. Trans. Amer. Math. Soc. **48**, 264–265 (1975)
107. Newman, D.J.: Simple analytic proof of the prime number theorem. Amer. Math. Monthly. **87**, 693–696 (1980)
108. Nielsen, N.: Handbuch der Theorie der Gammafunktion. Teubner, Leipzig (1906) Chelsea, New York (1965)
109. Noble, B.: Methods Based on the Wiener-Hopf Technique. Chelsea, New York (1988)
110. Ohsawa, T.: Analysis of Several Complex Variables. American Mathematical Society, Providence (2002)
111. Olver, F.W.J., Lozier, D.W., Boisvert, R.F., Clark, C.W.: NIST Handbook of Mathematical Functions. Cambridge University Press, Cambridge (2010)
112. Osgood, B.G.: Lectures on the Fourier Transform and its Applications. Amer. Math. Soc, Providence (2019)
113. Ovsienko, V., Tabachnikov, S.: Projective Differential Geometry. Old and New. Cambridge University Press, Cambridge (2005)
114. Ovsienko, V., Tabachnikov, S.: What is the Schwarzian derivative. Notices of the A. M. S. **56**, 34–36 (2009)
115. Paley, R.E.A.C., Wiener, N.: Fourier Transforms in the Complex Domain. Amer. Math. Soc., Providence (1934), reprint 1987
116. Pleijel, A.: On a theorem by P. Malliavin. Israel J. Math. **1**, 166–168 (1963)
117. Protter, M.H., Weinberger, H.F.: Maximum Principles in Differential Equations. Springer, New York, Heidelberg, Berlin (1984)
118. Pucci, P., Serrin, J.: The Maximum Principle. Birkhäuser, Basel (2007)
119. Remmert, R.: Wielandt's theorem about the $\Gamma$-function. Amer. Math. Monthly **103**, 214–220 (1996)
120. Riemann, B.: Ueber die Anzahl der Primzahlen unter einen gegebenen Grösse. Monatsber. der Berliner Akad. Nov. (1859)
121. Roy, R.: Sources in the Development of Mathematics. Cambridge University Press, Cambridge (2011)
122. Roy, R.: Elliptic and Modular Functions from Gauss to Riemann. Cambridge University Press, Cambridge (2017)
123. Ru, M.: Nevanlinna Theory and its Relation to Diophantine Approximation. World Scientific, River Edge N. J (2001)
124. Rubel, L.: Entire and Meromorphic Functions. Springer, New York, Heidelberg, Berlin (1996)
125. Schlag, W.: A Course in Complex Analysis and Riemann Surfaces. Amer. Math. Soc, Providence (2014)
126. Siegel, C.L.: A simple proof of $\eta(-1/\tau) = \eta(\tau)\sqrt{\tau/i}$. Mathematika **1**, 4 (1954)

127. Siegel, C.L.: Topics in Complex Function Theory, vol. I. Wiley-Interscience, New York (1969)
128. Siegel, C.L.: Topics in Complex Function Theory, vol. II. Wiley-Interscience, New York (1971)
129. Stahl, S.: A Gateway to Modern Geometry. The Poincaré Half-Plane. Jones and Bartlett, Boston (1993)
130. Stein, E.M.: Interpolation of linear operators. Trans. Amer. Math. Soc. **83**, 482–492 (1956)
131. Stein, E.M.: Singular Integrals and the Differentiability Properties of Functions. Princeton University Press, Princeton (1970)
132. Stein, E.M., Shakarchi, R.: Complex Analysis. Princeton University Press, Princeton (2003)
133. Tauber, A.: Ein Satz aus der Theorie der unendlichen Reihen. Monatsh. Math. Phys. **8**, 272–277 (1897)
134. Titchmarsh, E.C.: The Theory of Functions. 2nd edn. Oxford University Press, Oxford (1939), reprint 1958
135. Titchmarsh, E.C.: The Theory of the Riemann Zeta Function, 2nd edn. Cambridge University Press, Cambridge (1986)
136. Titchmarsh, E.C.: Introduction to the Theory of Fourier Integrals, 3rd edn. Chelsea, N.Y. (1986)
137. Venkov, A.B.: Spectral Theory of Automorphic Functions and Its Applications. Kluwer, Dordrecht (1990)
138. Walker, P.: Elliptic Functions. A Constructive Approach. Wiley, Chichester (1996)
139. Weil, A.: Elliptic Functions According to Eisenstein and Kronecker. Springer, New York, Heidelberg, Berlin (1976), reprint 1999
140. Weyl, H.: The Concept of a Riemann Surface. Addison-Wesley, Reading, Mass (1964)
141. Widom, H.: Wiener-Hopf integral equations. In: The Legacy of Norbert Wiener: A Centennial Symposium. Amer. Math. Soc, Providence (1994)
142. Wiener, N.: Tauberian theorems. Ann. Math. **33**, 1–100 (1932)
143. Wiener, N.: The Fourier Integral and Some of its Applications. Cambridge University Press, Cambridge: reprint 1988, p. 1959. Dover, New York (1933)
144. Wiener, N., Hopf, H.: Über eine Klasse singulárer Integralgleichungen. S.B. Preuss. Akad. Wiss. Berlin. Phys. Math. Kl.**30/32**, 696–706 (1931)
145. Wong, R.S.C.: Asymptotic Approximations of Integrals. Soc. Ind. Appl. Math, Philadelphia (2001)
146. Yang, L.: Value Distribution Theory. Springer, New York, Heidelberg, Berlin (1993)
147. Zagier, D.: Newman's short proof of the prime number theorem. Amer. Math. Monthly **104**, 705–708 (1997)
148. Zheng, J.: Value Distribution of Meromorphic Functions. Tsinghua University Press, Beijing (2010)
149. Zygmund, A.: Trigonometric Series, vol. I, II, 3rd edn. Cambridge University Press, Cambridge (2002)

# Index





R. Beals and R. S. C. Wong, *Explorations in Complex Functions*,
Graduate Texts in Mathematics 287,
https://doi.org/10.1007/978-3-030-54533-8